(12)

NPRDC SR 83-2

OCTOBER 1982

HANDBOOK FOR TESTING IN NAVY SCHOOLS

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
SAN DIEGO, CALIFORNIA

NAVY PERSONNEL RESEARCH
AND
DEVELOPMENT CENTER
San Diego, California 92152

NPRDC Special Report 83-2

October 1982

HANDBOOK FOR TESTING IN NAVY SCHOOLS

John A. Ellis
Wallace H. Wulfeck, II

Reviewed by
John D. Ford, Jr.

Released by
James F. Kelly, Jr.
Commanding Officer

Navy Personnel Research and Development Center
San Diego, California 92152

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
NPRDC SR 83-2

2. GOVT ACCESSION NO.
AD-A122 479

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
HANDBOOK FOR TESTING IN NAVY SCHOOLS

5. TYPE OF REPORT & PERIOD COVERED
Special Report

6. PERFORMING ORG. REPORT NUMBER
13-80-11

7. AUTHOR(s)
John A. Ellis
Wallace H. Wulfeck, II

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Navy Personnel Research and Development Center
San Diego, California 92152

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
PE 63720N
Z1175.PN-05

11. CONTROLLING OFFICE NAME AND ADDRESS
Navy Personnel Research and Development Center
San Diego, California 92152

12. REPORT DATE
October 1982

13. NUMBER OF PAGES
171

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Criterion-referenced testing
Test item construction
Test item analysis

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This handbook provided by this report provides detailed "how-to" procedures for use by course developers in constructing test items and tests for Navy technical courses.

FOREWORD

This handbook was developed under program element 63720N, project Z1175.PN (Training System Design and Management), subproject 05 (Improved Effectiveness in Course Design, Delivery, and Evaluation) and was sponsored by the Deputy Chief of Naval Operations (Manpower, Personnel, and Training) (OP-01). The objective of the subproject is to develop an empirically-based instructional design support system to aid developers in choosing instructional alternatives based on costs/benefits and specific resource limitations. The purpose of this handbook is to provide detailed "how-to" procedures for use by course developers in constructing test items and tests in Navy technical courses. Funds for its final development and tryout were provided by the Chief of Naval Education and Training.

The handbook is based on work covered in a number of previous Center reports; specifically, those describing the instructional quality inventory (NPRDC Spec. Reps. 79-3, 79-5, 79-24, and 80-25) and criterion-referenced tests (NPRDC Spec. Rep. 80-15 and Tech. Notes 80-8 and 81-6). It is intended for use by test designers and developers in the Naval Education and Training Command. A workbook, for use in conjunction with this handbook, is provided by NPRDC SR 83-7.

JAMES F. KELLY, JR.
Commanding Officer

JAMES W. TWEEDDALE
Technical Director

SUMMARY

Problem

Modern military instruction is developed according to a systematic method called Instructional Systems Development (ISD). This method includes the following steps:

1. Job/task analysis leading to specification of training objectives.

2. Development of tests to measure student progress toward the objectives or to diagnose areas of student weakness.

3. Design of new instruction and/or adaptation of existing instruction to achieve the objectives.

4. Implementation of the training program.

5. Evaluation and feedback for course maintenance.

ISD is documented in NAVEDTRA 106A, Interservice Procedures for Instructional Systems Development (CNET, 1975), and in NAVEDTRA 110A, Procedures for Instructional Systems Development (CNET, 1981).

Testing of student performance is an important part of ISD, since the adequacy and maintenance of any training program depends on careful assessment of the quality of student learning.

Purpose

The Handbook for Testing in Navy Schools is intended to supplement NAVEDTRA 106A and 110A by providing "how-to-do-it" procedures for the development, implementation, and evaluation of test items and tests. It is intended for use by test designers and developers in the Naval Education and Training Command.

Overview

Chapter 1 of this handbook provides an introduction to testing in Navy technical training and the process of test item and test development. Chapters 2 through 5 contain detailed procedures for constructing different types of test items and tests to measure different types of student behavior and to diagnose in detail reasons for inadequate student performance. The last chapter provides procedures for pilot-checking tests, together with statistical procedures to be used for evaluating test items and tests.

CONTENTS

CHAPTER 1

INTRODUCTION TO TESTING

Purposes of Testing

Tests are used to make decisions about people. These decisions determine the types of tests to be used or the way test results will be interpreted. There are two main types of decisions:

1. How does an individual's performance compare with the performance of others?

2. How does an individual's performance compare with a preestablished standard of performance?

The first type of decision involves comparing people with others. For example, it might be necessary to select people for advanced training, rank people relative to each other, or decide who should be promoted. Tests used to help make those decisions include tests like college entrance examinations, advancement-in-rate exams, most IQ tests, most "standardized" tests used in public schools, the Armed Services Vocational Aptitude Battery (ASVAB) tests, and many others. Tests of this type are called NORM-REFERENCED TESTS. "Norm-referenced" means that an individual's test performance is compared--referenced--to the average--norm--of everyone else's test performance. In norm-referenced tests, the individual's raw score (the actual number of items correct) cannot be interpreted without knowing how other people scored. For example, an individual who got 50 items right out of 100 might have done very well on a test if everyone else scored lower. For this reason, raw scores are usually transformed (changed) into "standard" scores (like percentiles) so that individuals can be compared.

The second type of decision involves comparing a person's performance with an absolute standard. The performance of other individuals is irrelevant. For example, it might be necessary to decide whether an individual should or should not be allowed to drive a car, be licensed to practice medicine, become a building contractor or real-estate broker, be graduated from a school, pass from one lesson to another, or be put in a remedial program. Tests of this type are called CRITERION-REFERENCED TESTS. "Criterion-referenced" means that an individual's test performance is compared--referenced--to an absolute standard or criterion concerning what a person must know or be able to do. Therefore, in criterion-referenced tests, an individual's raw score--the actual performance--is meaningful; the individual either meets the criterion or does not.

It is important to realize that a test is criterion-referenced or norm-referenced, depending on the interpretation of scores--not necessarily on the content of the test. It is often difficult to tell if a test is criterion-referenced or norm-referenced just by looking at the test items. The scoring, interpretation, and decisions made on the basis of test results determine the type of test. In norm-referenced tests, a single percentage score is usually reported, which tells how the student compares with other people who took the test. This score does NOT provide information about specific knowledge or performance ability. Criterion-referenced tests, on the other hand, provide specific information about what a student can or cannot do on specific objectives. In criterion-referenced tests, it does not usually make sense to give a "number" score by itself. Instead, criterion-referenced tests are interpreted in terms of particular skills or knowledge a student may or may not have.

As discussed above, criterion-referenced tests are used to determine whether an individual can or cannot perform relative to an explicit criterion. In many cases, when an individual cannot perform, it is important to determine in more detail the reasons for inadequate performance. In this situation, a special type of criterion-referenced test, called a DIAGNOSTIC TEST, is used. Diagnostic tests are criterion-referenced tests but are designed to give information about what an individual does not know or can not do and why. Diagnostic tests are not appropriate for all types of tasks; some tasks are so simple and straightforward that the criterion test is all that is necessary. Other tasks, however, are complicated enough so that there may be many reasons for performance failure. Diagnostic tests are used when it is necessary to get more information about the reasons for performance failure. Then, one can make more precise decisions about what remedial training should be provided or whether or not additional training is necessary.

Testing in the Navy

Both criterion-referenced and norm-referenced tests are used in the Navy. Norm-referenced tests are used for initial selection and assignment (the ASVAB) and for advancement-in-rate decisions. All other tests used in the Navy, particularly in technical training, are (or should be) criterion-referenced.

In Navy schools, there may be pretests, progress tests (for units, lessons, or modules), comprehensive within-course tests, and final comprehensive tests. These may take the form of written tests, job-performance tests, laboratory exercises, etc. All of these are meant to determine whether or not the student has sufficient knowledge or skill to meet criterion; that is, whether or not the student has completely learned the material and, if not, why. The philosophy underlying Navy testing is based on the achievement of learning objectives. Tests are given to determine whether a student has reached the criterion specified in an objective. If a student performs poorly on a test, this means that there is some material the student has not learned. Remediation must be given and, in many cases, the student must be retested until the criterion is reached.

Remediation and Testing

Remediation is an important part of the instructional process leading to achievement of objectives. It is important because it is unrealistic to expect all students to pass all tests the first time. Therefore, remediation must be planned for in the instructional and test development process. There are a number of ways in which students can be remediated:

1. Informal discussion with the instructor about test items missed. This can be done on a one-to-one basis in an individualized course. In a group paced course, the instructor can review with the class all the test items or objectives if item security is necessary.

2. Written feedback on the test.

3. An assignment by the instructor (or test scorer) to review the instructional materials relevant to the missed items.

4. An assignment by the instructor (or test scorer) to study additional remedial materials relevant to the missed items. If remedial materials are to be included in the course, they should be developed at the same time the course materials are developed and specific instructions for their use should be included in the curriculum outline and course delivery plan.

Retesting can be done wi n any of these remediation methods and should be done to ensure that the student has achieved the objectives. However, there are some cases in which retesting is unnecessary. These will be discussed in a later section.

In summary, this "test-for-achievement-of-learning-objectives, remediate, retest when necessary" process is the basis for this handbook. This handbook is therefore concerned only with criterion-referenced testing--not with norm-referenced testing.

Overview of Test Development

Technical training programs in the Navy are developed according to the Interservice Procedures for Instructional Systems Development (ISD), as described in NAVEDTRA 106A and 110 or MIL-STD 1379B. The ISD process requires that test items be written from learning objectives. This development process involves the following major steps:

1. Job/task analysis and specification of learning objectives. Before any test development can be done, analysis must be conducted to determine (a) the tasks to be trained and tested and (b) the knowledge and skill requirements for these tasks. The tasks, knowledges, and skills are then further analyzed to determine actual learning objectives for the training program. These objectives are the basis for test development, since they specify the criteria for measuring student achievement. This handbook assumes that analysis and objectives specification have been completed.

2. Determine test item type, format, diagnostic requirements, and scoring standards on the basis of objectives. Well-stated learning objectives serve as design guides for test items. They specify the conditions under which the student will perform, the behavior that must be exhibited during testing, and the standards that must be met. This information also helps determine appropriate test item type, format, and diagnostic requirements. In some cases, however, fiscal, time, and scoring constraints place limits on the testing options available. In these cases, the constraints and the limits they place on testing the objective should be noted on the Learning Objective Analysis Worksheet (LOAW, CNET-GEN 1550/4 (Rev 6-81).

3. Write test items. The test item design specifications from step 2 are used here to (a) write actual test items, (b) group them into lesson, module, final, etc., tests, (c) develop directions to students about how to take the tests, and (d) develop directions for test administrators about how to give and score the tests. There are a variety of prescriptions about how to develop good test items that must be kept in mind; these vary depending on the type and format of the test items.

4. Pilot checking of tests. The tests from step 3 are administered to representative samples of students and job experts. Results are analyzed to determine deficiencies in test items, and the items are revised accordingly.

5. Implementation, evaluation, and revision. Plans are developed for the operational administration of the tests, including scoring, test security, resource requirements, data collection, and data summarization. The testing program is then implemented. Data are used to maintain the quality of the program through ongoing evaluation and revision

Setting Standards for Criterion-referenced Tests

Because Navy training is concerned with the achievement of the objectives to be learned, setting standards for tests is very important. Unfortunately, many test developers pay little attention to determining standards for tests. Often they use general rules of thumb such as "three out of five" or "62.5, stay alive." The problem is that these standards are based neither on a logical analysis of the content area nor on job performance requirements.

In any training program in which students are required to achieve specific learning objectives, standards for testing must match the standards specified in the objectives. How can student competence be certified if students are not required to demonstrate achievement of objectives during tests? The logic of ISD requires that job performance requirements be reflected in training objectives and test items. Since there are few job performance requirements that allow for the job to be performed incompletely or inaccurately, tests must require the student to demonstrate achievement of learning objectives.

This logic also applies to "memory-level" knowledge information that is included in a training program because it supports job performance. The criterion here should also match the learning objective, because incomplete knowledge means that job performance will suffer.

There are a few Navy tasks where the standard may be difficult to determine, because the nature of the job makes it hard to define acceptable performance. For example, detection tasks often cannot be performed "perfectly," even by highly skilled experts. But here, it is the nature of the job that limits performance, not some skill or knowledge deficiency. Standards for tasks like this are developed by analyzing expert job performance or by interviewing job experts. There are also tasks that have rate or tolerance standards (like typing 45 words-per-minute with less than five errors). Here, students should achieve whatever is specified in the objective(s).

In summary, for all training tasks and related knowledge information, tests should require achievement of the standards specified in the objective(s).

Standards and Remediation Plans

On any particular test, it is unrealistic to expect all students to achieve the objective(s) on their first try. This does not mean that the test standard should be set lower; it simply means that students' weaknesses should be diagnosed, and remediation should be given so that the objective(s) is achieved. An important aspect of remediation is retesting. If the student is not retested, then you cannot be sure that the objective has been achieved. Without remediation and retesting, the standards established for the objectives have no meaning, because it makes no sense to give tests during training unless test results are used to improve student performance. The way to do this efficiently is to make tests as diagnostic as possible, so that remediation can be prescribed on the basis of test results. Standards should be set by consulting with subject matter experts. Guidelines for setting standards for different types of test items are given in the appropriate chapters of this handbook.

Organization of this Handbook

This handbook is organized around the main test development steps described above. As stated earlier, this handbook assumes that job/task analyses have been completed and training objectives have been specified. Chapter 2 deals with classification and quality control of objectives. A scheme for classifying objectives is presented, and the implications of the classification system for test item design and test development are discussed.

Chapters 3, 4, and 5 deal with the design and development of test items for different types of objectives. Chapter 3 concerns test items for objectives that primarily require recall of information. Chapter 4 is concerned with testing procedural skills. Chapter 5 deals with tests for objectives that require transfer or generalization of knowledge or skill. Finally, Chapter 6 deals with pilot checking of tests and includes methods for using statistical analyses to aid in quality control of test items.

CHAPTER 2

CLASSIFICATION OF OBJECTIVES AND IMPLICATIONS FOR TESTING

Introduction

Before we can start developing test items, we need to talk about designing test item specifications that ensure appropriate measurement of learning objective achievement. The learning objectives for a course serve as the basis for specifying test item requirements. Different course objectives, however, may have widely differing testing requirements because there are many different types of behavior that objectives can require. To make test item specification more precise, we have worked out a fairly simple procedure for classifying learning objectives according to the nature and type of behavior required. This handbook is organized so that, once an objective has been classified, you can use the procedures in the appropriate chapter in the handbook to develop the appropriate kind and amount of test items needed to measure achievement of that objective adequately.

Classification of objectives is necessary for several reasons. First, it helps make more precise judgments about the adequacy of learning objectives and leads to more precise test item specifications. The classification scheme described in this chapter was designed to have direct implications for how test items for an objective should be developed. The classification scheme can also be used to evaluate objectives and test items that already exist. It was designed so that judgments can be made about how consistent objectives and test items are with each other. If we didn't classify objectives and test items, all we could say is, "This is an objective and this is a test item, and they don't look too different." The classification scheme allows us to ensure that an objective and its corresponding test item both address the same thing. The classification scheme also helps us judge whether or not objectives and test items are adequate. The scheme presented in this chapter was adopted from the Instructional Quality Inventory (IQI).*

*The IQI is described in four volumes:

Vol. I. Introduction and Overview	(NPRDC SR 79-3)
Vol. II. User's Manual	(NPRDC SR 79-24)
Vol. III.Training Workbook	(NPRDC SR 80-25)
Vol. IV. Job Performance Aid	(NPRDC SR 79-5)

The Classification Scheme

Objectives and test items can be classified according to:

1. What the student must do; that is, the TASK to be performed.

2. The instructional CONTENT; that is, the type of information the student must learn.

The Task Dimension

A student can either REMEMBER information or USE the information to do something. This distinction corresponds to the difference between knowledge and application and between declarative and procedural knowledge. The following two test items illustrate the REMEMBER-USE distinction.

1. REMEMBER: The symbol for resistor is__________.

2. USE: Using your knowledge of electronic theory, predict what would happen in the circuit shown below if the load resistance were shorted?

These two test items differ with respect to what the student is supposed to do (Task). In the first item, the student has to REMEMBER something. In the second, he has to apply or USE his knowledge in a new situation.

The Content Dimension

There are five types of content: FACTS, CATEGORIES, PROCEDURES, RULES, and PRINCIPLES. FACTS are simple associations between names, objects, symbols, locations, etc. Facts can only be remembered, while the other content types can be remembered or used. CATEGORIES are classifications defined by certain specified characteristics. PROCEDURES consist of ordered sequences of steps or operations performed on a single object or in a specific situation. RULES also consist of ordered sequences of operations, but they can be performed on a variety of objects or in a variety of situations. PRINCIPLES involve explanations, predictions, or diagnoses based on theoretical or cause-effect relationships. The following examples of objectives and test items illustrate the five content areas for the REMEMBER task level.

1. REMEMBER FACT.

   a. The symbol for resistor is____________. (test item)

   b. The student will list the names of the parts of the distributor. (objective)

2. REMEMBER CATEGORY.

   a. List the defining characteristics of a jet pump. (test item)

   b. The student will define the various kinds of clouds (cumulus, stratus, etc.). (objective)

3. REMEMBER PROCEDURE.

   a. List, in order, the steps for dissassembling a carburetor. (test item)

   b. The student will describe the procedure for preparing and sending a radio message. (objective)

4. REMEMBER RULE.

   a. List the steps involved in finding the rhumb-line course between two points on the earth. (test item)

   b. The student will state the general rule for solving the circuit current, given voltage and resistance. (objective)

5. REMEMBER PRINCIPLE.

   a. State the principles of electron movement in a semiconductor junction. (test item)

   b. The student will recall the reasons why hydraulic fluid contamination must be avoided. (objective)

FACTS can only be remembered but, for the other content types, the student may be asked to use his knowledge to classifv, perform, solve, or predict. The following are examples of the USE task level for all content types except facts.

1. USE CATEGORY.

   a. Which of the clouds pictured below are cumulus clouds? (test item)

   b. Given photographs of facial expressions, the student will classify them as to emotional type. (objective)

2. USE PROCEDURE.

   a. Clean an M-16 rifle. (test item)

   b. The student will make and bake a cherry pie. (objective)

3. USE RULE.

   a. Calculate the rhumb-line course from Pearl Harbor to Long Beach, CA. (test item)

   b. Given the values for voltage and resistance, the student will calculate the current flow. (objective)

4. USE PRINCIPLE.

   a. Predict what would happen in the circuit shown below if the load resistance were shorted. (test item)

   b. The student will predict what is likely to occur if the landing gear fluid were contaminated with sand. (objective)

The USE-level can be further divided into two types: (1) USE-UNAIDED, where the student has no aids except his own memory, and (2) USE-AIDED, where the student has a job aid to perform the task. For the USE-AIDED type, the nature of the aid depends on the content type. For USE-AIDED CATEGORY, the aid consists of a decision strategy including each critical characteristic. In simple cases, when the aid may comprise only a list of characteristics, the decision strategy would be implied. For USE-AIDED PROCEDURES, the aid is a list of steps to be performed. For USE-AIDED RULES, the aid is at least a statement of the formula or rule to be applied and could include guidelines for when and how to apply it. For USE-AIDED PRINCIPLES, the aid is also at least a statement of the principle and could include guidelines for when and how to apply it. In summary, the REMEMBER-level involves "pure" remembering; the USE-UNAIDED-level, remembering what is to be used and then using it; and the USE-AIDED-level, "pure" using.

The Classification Process

Classifying objectives, test items, and instruction is a simple two step process. First, you determine whether the task level is REMEMBER, USE-UNAIDED, or USE-AIDED. Second, you determine whether the content type is FACT, CATEGORY, PROCEDURE, RULE, or, PRINCIPLE. These steps are explained below.

Step 1. Determine the TASK LEVEL.

a. Determine whether the student is to REMEMBER or USE information.

b. If the student is to USE information, determine whether the task level is USE-AIDED or USE-UNAIDED.

Step 2. Determine the CONTENT TYPE.

a. If the student must recall or recognize names, parts, locations, functions, dates, places, etc., the content type is FACT.

b. If the student must remember characteristics of similar objects, events, or ideas according to characteristics, or if the student must sort, classify, or categorize objects according to characteristics, the content type is CATEGORY.

c. If the student must remember a sequence of steps that apply to a single situation, or if the student must actually perform the sequence of steps, the content type is PROCEDURE.

d. If the student must remember a sequence of steps and decisions that apply in a variety of situations, or if the student must apply the sequence across a variety of situations or types of equipment, the content type is RULE.

e. If the student must remember how or why things work the way they do, or cause-effect relationships, or if the student must explain how things work, predict what will happen given a set of boundary conditions, or diagnose causes given a set of symptoms, the content type is PRINCIPLE.

This procedure is explained in more detail on the following pages.

Explanation for Step 1, Determine Task Level

The first step in the classification procedure involves deciding whether the student is required to REMEMBER information, or whether the student must USE information to perform some task. The REMEMBER-USE distinction is a simple one. The determination can usually be made by looking at the action in the objective or test item. Typical action verbs are listed below. The ones on the left usually indicate REMEMBER-level tasks, while the ones on the right usually indicate USE-level tasks.

REMEMBER

name
state (from memory)
list (from memory)
recall
remember
relate
write (from memory)
recognize
explain (from memory)
describe (from memory)

USE

apply
classify
analyze
derive
demonstrate
discriminate
evaluate
solve
prove
sort
explain
maintain
compute
determine
.
.
. (there are many other USE actions.)

operate
repair
adjust
calibrate
remove
replace
assemble
disassemble
calculate
troubleshoot
load
unload
predict

If the task level is USE, the next step is to determine whether an aid is given. This can be done by looking at the "conditions" part of the objective or test or practice item. Anything that replaces the need for memory counts as an aid.

AIDS include:

1. A list of procedure steps from a technical manual or MRC card.

2. A formula for solving problems.

3. A list or table or chart of characteristics.

4. A statement of a principle.

Normal tools, materials, etc., are NOT aids.

Explanation for STEP 2, Determine Content Type.

STEP 2a: FACTS. Facts are what you think they are. They are simple associations between objects, events, names, parts, functions, locations, dates, etc. Facts don't have to come in pairs; there may be three or four or more pieces of information that go together. For example, a student might have to remember the name, location, and function of each of the parts in some piece of equipment.

Key words or phrases that may help identify FACT-level objectives or test items are listed below.

The student will give the symbol for each ....

match each ... with its ....

list the names of each ....

recall the dates of ....

recall the location and function of each ....

give the ... associated with each ....

STEP 2b: CATEGORIES. Categories refer to groups of similar objects, events, or ideas. They are similar, or are grouped together, because they have characteristics in common. Category tasks nearly always involve classification or sorting on the basis of these critical characteristics. At the REMEMBER-level for categories, the student is required to remember these characteristics and how they go together. At the USE-level, the student is required to identify, sort, or classify things according to these characteristics.

Key words or phases that may help identify CATEGORY-level objectives or test items are listed below.

1. REMEMBER-level.

	recall		characteristics		type of ...
	list		features		kind of ...
The student will	name	the	definition	of each	category of ...
	describe		attributes		classification
	give				situation

2. USE-level.

	sort			type
	classify			kind
The student will	categorize	each ...	according to	characteristics
	identify			definition
	recognize			features
	choose			
	select			

STEP 2c: PROCEDURES. A procedure is a sequence of steps that must be performed in order. Procedures are always applied, in the same way, on situations or equipment that do not change. At the REMEMBER-level for procedures, the student is required to remember the steps and their order. At the USE-level, the student is to perform the procedure. Key words or phrases are listed below.

1. REMEMBER-level.

	recall		steps		operating
	list		process		performing
The student will	name	the	procedure	for	maintaining
	state		sequence		lighting off
	give				etc.

2. USE-level.

	apply	remove
	operate	replace
The student will	repair	assemble
	adjust	produce
	calibrate	destroy

etc.

STEP 2d: RULES. A rule, like a procedure, is a sequence of steps. However, rules can be applied in a variety of situations or on a variety of different equipments. Because they apply in a variety of situations, rules sometimes have complicated decision steps.

Formulas and mathematical calculations always involve the use of rules, unless the student has a calculator or computer. Key words or phrases are listed below.

1. REMEMBER-level.

	recall		formula		solving
	name		rule		deriving
The student will	state	the	law	for	proving
	give		process		calculating
	remember		steps		determining
					etc.

2. USE-level.

	solve	find
	derive	translate
The student will	prove	program
	calculate	add
	determine	subtract

etc.

STEP 2e: PRINCIPLES. Principles involve explanations of why or how things work the way they do, or predictions about "what would happen if ....," or diagnosis of why or how something happened. Principles are based on cause-effect relationships, theoretical statements, statistical associations, or physical or scientific "laws." At the REMEMBER-level, the students must recall reasons, causes and effects, theoretical statements, etc. At the USE-level, students must use their knowledge to give an explanation about how something works, or predict what is likely to happen, or diagnose why something isn't working the way it should. Key words or phrases are listed below.

1. REMEMBER-level.

	recall	the principle of
	remember	the explanation of
The student will	state	how ...
	describe	why ...
	discuss	the reasons for

2. USE-level.

	analyze
	evaluate
The student will	explain
	diagnose
	troubleshoot
	predict

Note. It is sometimes difficult to determine whether a PRINCIPLE objective is REMEMBER or USE. Look at the following example objective:

> "The student will explain the operation of a rotary-gear pump."

This objective could be either USE or REMEMBER. It would be USE if the student had not been taught about rotary-gear pumps and was required to use his knowledge of hydraulic principles to explain their operation. If the instruction had dealt with rotary gear pumps specifically, then the objective would be REMEMBER.

If you encounter an objective of this type, you must determine if the information has been taught (if you are evaluating an existing course) or if the information is to be taught (if you are evaluating the objectives of a course under development). If the information is or will be included in the instruction, the task level is REMEMBER. If the student is required to use principles to explain things that have not been taught specifically, the task level is USE.

Actually, it is best to make the objective precise in the first place. Example:

> "The student will recall the explanation of the operation of a rotary-gear pump."

Additional Explanation for STEP 2, Determine Content Type

On the preceding pages, we have given definitions and key words for each content type; however, determining content type can still be difficult. In the following sections, we will give a schematic representation of each content type, and further guidelines for distinguishing among content types.

FACTS. Facts are simple associations between names, objects, etc. The task is always for the student to recall them or, given one part of the fact, to recall the other parts.

fact 1: 0---0---0---0

fact 2: 0---0---0---0

.

.

.

fact n: 0---0---0---0

Example:

"The student will recall in writing the name, location, and function of each dial on the front panel of ...."

Dial 1: name--location--function

Dial 2: name--location--function

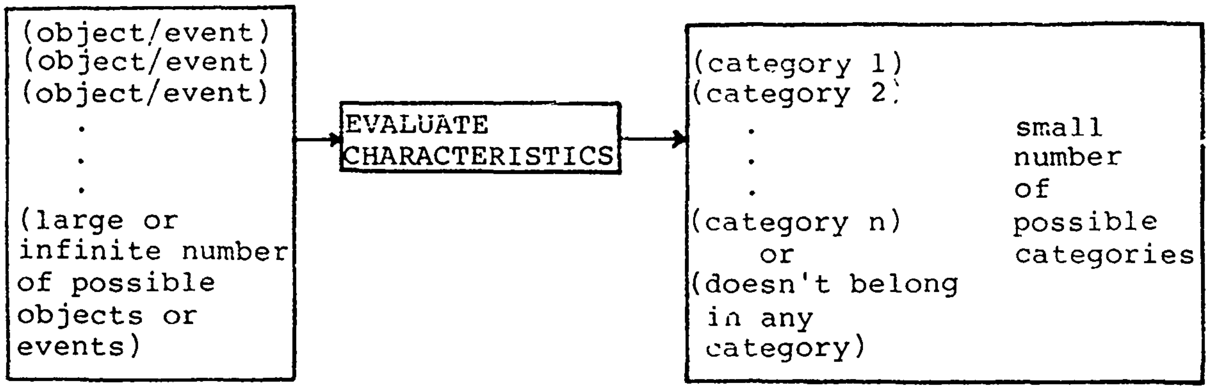
.

.

.

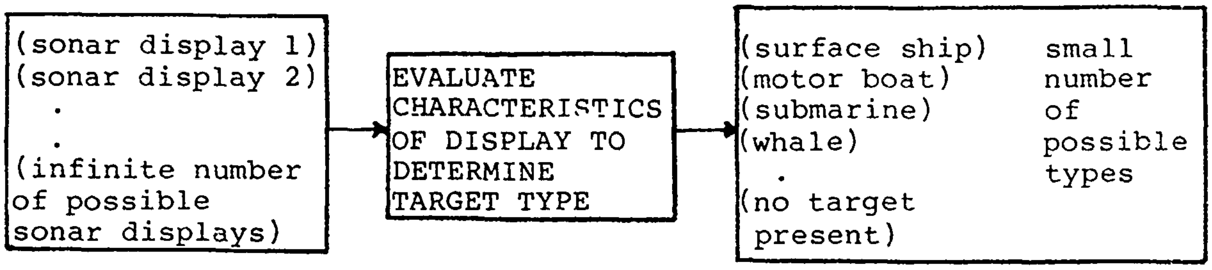
Dial n: name--location--function

CATEGORIES. Category tasks involve sorting or classifying objects or events according to their characteristics or features. At the REMEMBER-level, the student must remember the characteristics and how they go together. At the USE-level, the student must identify, sort, or classify things according to these characteristics.


(object/event)
(object/event)
(object/event)
.
.
.
(large or infinite number of possible objects or events)
EVALUATE CHARACTERISTICS
(category 1)
(category 2)
.
.
.
(category n)
or
(doesn't belong in any category)
small number of possible categories

Example:

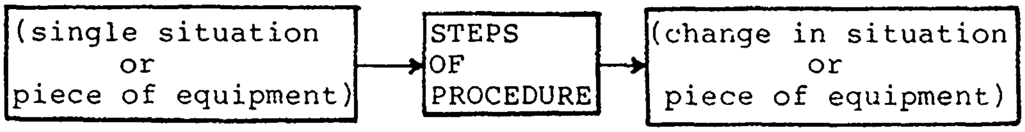
"Given a series of sonar scope displays, the student will classify them according to type of target."


(sonar display 1)
(sonar display 2)
.
.
(infinite number of possible sonar displays)
EVALUATE CHARACTERISTICS OF DISPLAY TO DETERMINE TARGET TYPE
(surface ship)
(motor boat)
(submarine)
(whale)
.
(no target present)
small number of possible types

For categories at the REMEMBER-LEVEL, the student must remember the characteristics. (The middle box above.)

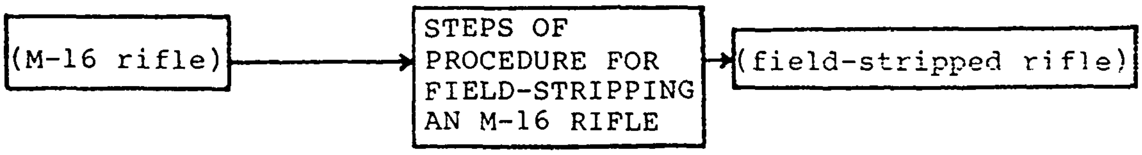
For categories at the USE-LEVEL, the student gets "inputs," evaluates the characteristics, and determines the appropriate category or type.

PROCEDURES. A procedure is a sequence of steps, performed in order, on a single piece of equipment or in a single situation. At the REMEMBER-level, the student must remember the steps in order. At the USE-level, the student is given a piece of equipment or a situation and must perform the steps.

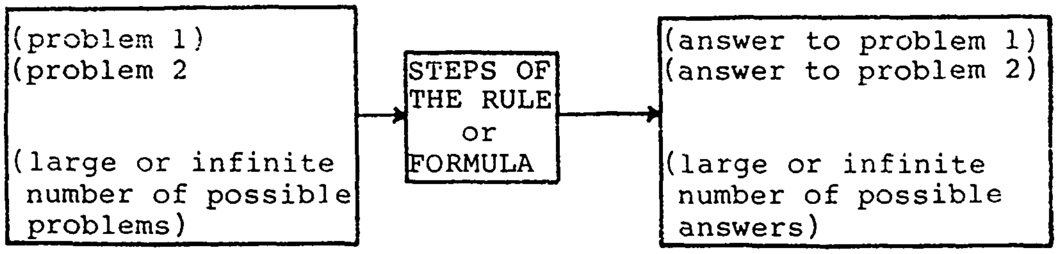

(single situation or piece of equipment)
STEPS OF PROCEDURE
(change in situation or piece of equipment)

Example:

"The student will field-strip an M-16 rifle."

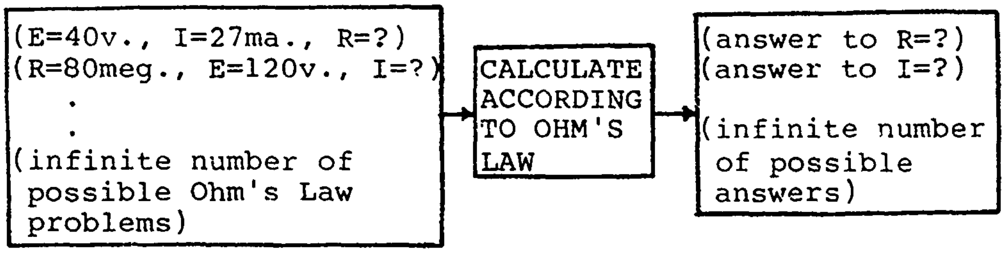


RULES. A rule, like a procedure, is a sequence of steps and decisions. However, rules can be applied in a variety of situations or on a variety of equipments. At the REMEMBER-level, the student must remember the steps and decisions. At the USE-level, the student is given problem situations and must apply the steps of the rule to solve the problem or come up with the answer.

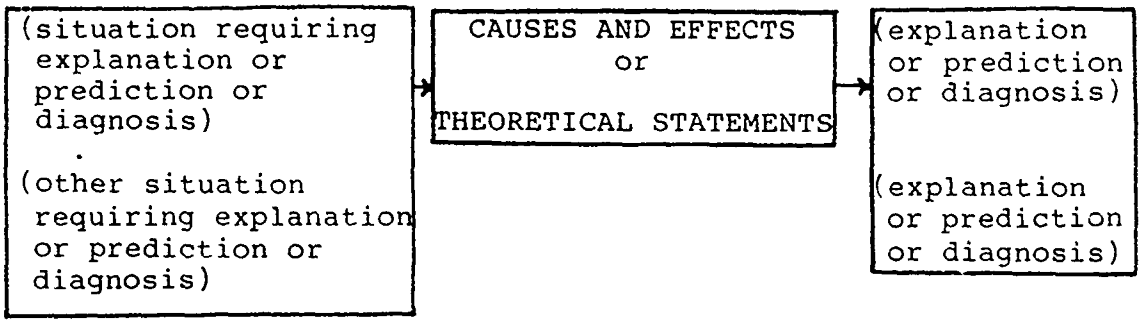

(problem 1)
(problem 2
(large or infinite number of possible problems)
STEPS OF THE RULE or FORMULA
(answer to problem 1)
(answer to problem 2)
(large or infinite number of possible answers)

Example:

"Given any two values of current, voltage, or resistance in a circuit, the student will use Ohm's Law to solve for the third value."

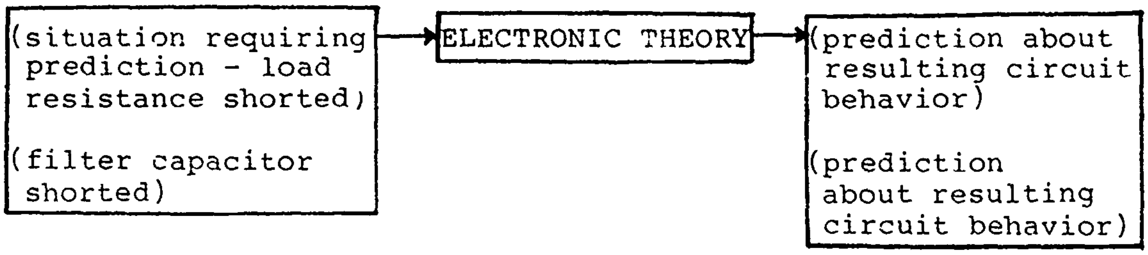

(E=40v., I=27ma., R=?)
(R=80meg., E=120v., I=?)
.
.
(infinite number of possible Ohm's Law problems)
CALCULATE ACCORDING TO OHM'S LAW
(answer to R=?)
(answer to I=?)
(infinite number of possible answers)

PRINCIPLES. Principles involve explanations of why or how things work the way they do, predictions about "what would happen if ...," diagnosis of why or how something happened, or why something doesn't work. At the REMEMBER-level, the student must recall reasons, causes and effects,theoretical statements, etc. At the USE-level, the student uses knowledge of causes and effects, or theories, to explain, predict, or diagnose.


(situation requiring explanation or prediction or diagnosis)
•
(other situation requiring explanation or prediction or diagnosis)
CAUSES AND EFFECTS
or
THEORETICAL STATEMENTS
(explanation or prediction or diagnosis)
(explanation or prediction or diagnosis)

Example:

"Based on knowledge of electronic theory, the student will predict the effect in the circuit shown below if the load resistance, or the filter capacitor, were shorted."


(situation requiring prediction - load resistance shorted)
(filter capacitor shorted)
ELECTRONIC THEORY
(prediction about resulting circuit behavior)
(prediction about resulting circuit behavior)

Note. The reason that principles are taught is that they apply in a variety of situations. They allow the student to make a variety of "what would happen if..." predictions or "what happened to..." diagnoses.

However, if the student is only required to remember one cause and one effect, then it should be classified as a fact. For example, if a particular symptom in a piece of electronic equipment always means that a particular part is damaged, that's a fact.

For CATEGORIES, PROCEDURES, RULES, and PRINCIPLES at the REMEMBER-level, the student has to remember whatever is in the middle box on the diagrams above. At the USE-level, the student has to perform the whole task.

The classification scheme can represented as a matrix with TASK LEVEL and CONTENT TYPE as the dimensions. This Task/Content Matrix is shown below.

	FACT	CATEGORY	PROCEDURE	RULE	PRINCIPLE
REMEMBER	RECALL OR RECOGNIZE NAMES, PARTS, DATES, PLACES, VOCABULARY DEFINITIONS, ETC.	REMEMBER THE CHARACTERISTICS OF EACH CATEGORY AND THE GUIDELINES FOR CLASSIFICATION.	REMEMBER THE STEPS OF THE PROCEDURE.	REMEMBER THE FORMULA OR THE STEPS OF THE RULE.	REMEMBER THE CAUSE AND EFFECT RELATIONSHIPS OR THE STATEMENT OF THE PRINCIPLE.
USE UNAIDED		CLASSIFY OR CATEGORIZE OBJECTS, EVENTS, IDEAS, ACCORDING TO THEIR CHARACTERISTICS, WITH NO MEMORY AID.	APPLY THE STEPS OF THE PROCEDURE IN A SINGLE SITUATION OR ON A SINGLE PIECE OF EQUIPMENT, WITH NO MEMORY AID.	APPLY THE FORMULA OR RULE TO A VARIETY OF PROBLEMS OR SITUATIONS, WITH NO MEMORY AID.	USE THE PRINCIPLE TO EXPLAIN, PREDICT, OR DIAGNOSE WHY OR HOW THINGS HAPPENED OR WILL HAPPEN, WITH NO MEMORY AID.
USE AIDED		GIVEN CATEGORY CHARACTERISTICS AND GUIDELINES, CATEGORIZE OBJECTS, EVENTS, IDEAS, ACCORDING TO CHARACTERISTICS.	GIVEN STEPS OF THE PROCEDURE, APPLY THE PROCEDURE IN A SINGLE SITUATION, OR ON A SINGLE PIECE OF EQUIPMENT.	GIVEN THE FORMULA OR RULE STEPS, APPLY THE FORMULA OR RULE TO A VARIETY OF PROBLEMS OR SITUATIONS.	GIVEN A STATEMENT OF THE PRINCIPLE, EXPLAIN, PREDICT, OR DIAGNOSE WHY OR HOW THINGS HAPPENED OR WILL HAPPEN.

Further Guidelines for Classification

Remember the Job. The most important thing to remember when attempting to classify objectives is the job. The classification scheme was designed so that classification depends on the job requirements. The most important requirement to consider is whether or not the student will have to deal with objects or situations that have not been seen or encountered during training. For the FACT and PROCEDURE content types, this does not occur. Facts by definition must be presented during training. The job requirements for procedures involve single pieces of equipment or single situations, and the student does not have to "generalize" to new equipments or situations. In other words, everything the student needs to know is presented during training.

On the other hand, there are some job situations that require the student to deal with so many possible objects, events, ideas, problems, or situations that it would be impossible to include all of them during training. In this case, the training program is designed so that the student will be able to deal with new cases. CATEGORIES, RULES, and PRINCIPLES are used in the classification scheme to cover this situation.

The CATEGORY content type is used when the job requires that a large number of possible objects, events, etc. be classified into, or identified as a member of, one of a small number of particular categories. Instead of having to remember each object and its classification, the student is given characteristics for each category, which allows classification of objects, etc., not seen before.

The RULE content type is used when the job requires that a large number of problems be solved or that a complicated sequence of steps be performed on a large number of different objects, events, etc. Instead of having to remember each problem or go through the steps on each object, the student is taught a RULE for dealing with problems, objects, and events not seen before.

The PRINCIPLE content type is used when the job requires prediction or interpretation of a large number of possible situations, events, effects, etc. Instead of having to remember each possible situation or event and its effects, the student is given a PRINCIPLE that summarizes the "how" or "why" of general situations or that allows the student to predict what is likely to occur in a variety of situations.

Problems in Classification. Sometimes classification can be tricky. There can be confusion between FACTS and CATEGORIES, and between PROCEDURES and RULES. The way to resolve problems is to "REMEMBER THE JOB"; that is, to consider carefully what the student must be able to do after instruction.

Again, the most important thing to consider is whether the student will have to deal with objects or situations that have not been seen during training. For example, if the student were required to sort or classify things according to their characteristics, and if the student on the job were going to be dealing with things not seen during training, then the objective would be a CATEGORY. However, suppose instead that there were only seven objects the student would ever see. Then it would be more efficient to teach each object and its category name as a fact (seven facts total).

Similarly, RULES are taught so that the student can apply knowledge to situations not seen in training. However, suppose the situations are so similar that "if you've seen one, you've seen them all." This would be more efficiently taught as a PROCEDURE.

On the other hand, some tasks look at first like PROCEDURES but turn out to be more complicated. An expert who really knows the job can help you make the decision.

Example: FACT vs. CATEGORY

"Given a variety of metal fasteners, the student will sort them according to type (bolts, screws, studs, or rivets)."

This could be taught as a CATEGORY: The student could be taught the characteristics of bolts (fine threads, blunt end, etc.), screws (course threads, pointed end, etc.), studs (no head, fine threads, etc.), and rivets (no threads, etc.). However, one bolt is pretty much the same as any other bolt, and the same for screws, studs, and rivets, except that they come in different sizes. Therefore, it might be more efficient to teach these as four FACTS: bolt - appearance, screw - appearance, etc. The confusion here can be solved if the job requirements are determined. If there are lots of different metal fasteners, and the student will see new bolts, etc. on the job, then the content type is CATEGORY. If there are only a few, and they're all nearly alike, then the content type is FACT.

Example: PROCEDURE vs. RULE

> "Given a word in print, correctly spelled, the student will look up the word in a dictionary, and state its definition orally."
>
> This might appear to be a RULE: There are a large number of possible words (inputs) and a large number of possible definitions (outputs). However, since the spelling is given, it's easy to look up the word: Find the first letter of the word, find that chapter in the dictionary, find the second letter, find that section of the chapter, etc. This is most efficiently taught as a PROCEDURE.
>
> However, suppose the word was given orally and not spelled. This would then be a fairly complicated RULE, involving listening skills, phonemic translations, etc.

Why do we need a classification scheme at all?

The classification scheme is essential for two reasons. First, it makes consistency judgments between objectives, test items, and instructional presentations possible. Second, it makes adequacy judgments about objectives, test items, and instructional presentations possible. In fact, the classification scheme was designed so that the classification of an objective has implications for the way the instruction for that objective should look. For example, the instruction for a USE-UNAIDED RULE should be different from the instruction for a REMEMBER-CATEGORY objective. The most important differences occur between the REMEMBER and USE TASK LEVELS, and between the CONTENT TYPES that do not require generalization (FACTS and PROCEDURES) and those that do (CATEGORIES, RULES, and PRINCIPLES). When generalization is required, there will be more examples and practice items covering a wider range of difficulty.

Volume II of the IQI series (Ellis, Wulfeck, and Fredericks, 1979) contains example objectives that are classified according to the process given here. Volume III (Fredericks, 1980) contains additional examples and classification practice.

Assessing the Quality of Objectives: The Objective Adequacy Procedures

As stated earlier, the training objectives for a course serve as the design specifications for test items and tests. Therefore, before test items are developed, the objectives must be checked to make sure that they are clearly stated and appropriate for the intent of the instructional program. Objectives are meant to communicate to everyone involved in an instructional program what the program is meant to accomplish; that is, what the students must be able to do upon completion, and, therefore, what they must be tested on. If an objective does not communicate the desired student performance clearly, or if it specifies something inappropriate for the intent of the course, then good test items cannot be developed from it.

For an objective to communicate clearly, it must specify three things. First, it must specify the CONDITIONS under which the student is to perform. Second, it must specify what STANDARDS the performance must meet. Third, it must specify what the performance is; that is, what ACTION the student is to perform. This information is the minimum needed; additional information might have to be provided to make the objective clear. Remember, the objective must communicate to test developers and to instructional developers. How could a test developer write an item if the standards were not known?

A good check on whether or not an objective is clear is to try to classify it according to the classification scheme in this chapter. If an objective is hard to classify (if it is hard to decide which box it goes in), this means that the ACTION is unclear; we don't know exactly what the student must do.

An objective may be clear, but be inappropriate for the intent of the instructional program. In this case, the objective is still inadequate. To be appropriate, an objective must prepare students for what they will be required to do or know following the instructional program. This following duty could be anything from job performance, to on-the-job training, to another formal follow-on school; these are all "jobs" after a training program. To determine appropriateness of an objective, the key is to "REMEMBER THE JOB."

Explanation for the Objective Adequacy Procedure

On the next two pages, OBJECTIVE ADEQUACY procedures from Volume IV of the IQI (Ellis and Wulfeck, 1978) are reproduced. These procedures correspond to the criteria discussed above. The steps included in these procedures are described on the following pages.

INSTRUCTIONAL
QUALITY
INVENTORY

CARD 1

OBJECTIVE ADEQUACY

STEP 1: ENTER the COURSE TITLE and OBJECTIVE NUMBER at the top of the form.

STEP 2: Determine whether or not the OBJECTIVE is CORRECTLY STATED.

2a: Are the CONDITIONS under which student performance is expected specified?

[illegible] PHYSICAL (weather, time of day, lighting, etc.)
SOCIAL (isolation, individual, team, audience, etc.)
PSYCHOLOGICAL (fatigue, stress, relaxed, etc.)

[illegible] GIVEN INFORMATION (scenario, formula, values, etc.)
CUES (signals for starting or stopping)
SPECIAL INSTRUCTIONS

[illegible] JOB AIDS (cards, charts, graphs, checklists, etc.)
EQUIPMENT, TOOLS
TECHNICAL MANUALS

2b: Are the STANDARDS which the student performance must meet specified?

[illegible] COMPLETENESS (how much of the task must be performed)
ACCURACY (how well must each task be performed)
TIME LIMIT (how much time is allowed)
RATE (how fast must task be done)

[illegible] COMPLETENESS (what must finished product contain)
QUALITY (what objective standard must product meet)
JUDGEMENT (what subjective opinions must product satisfy)

2c: Is the ACTION the student must perform specified?

Is an action verb used to specify what the student must do?

Is only one action stated in the objective?

STEP 3: Determine whether or not the OBJECTIVE is CLASSIFIABLE? Does the OBJECTIVE fit in one and only one cell of the table below?

	FACT	CATEGORY	PROCEDURE	RULE	PRINCIPLE
REMEMBER	RECALL OR RECOGNIZE NAMES, PARTS, DATES, PLACES, VOCABULARY DEFINITIONS, ETC.	REMEMBER THE CHARACTERISTICS OF EACH CATEGORY AND THE GUIDELINES FOR CLASSIFICATION.	REMEMBER THE STEPS OF THE PROCEDURE.	REMEMBER THE FORMULA OR THE STEPS OF THE RULE.	REMEMBER THE CAUSE AND EFFECT RELATIONSHIPS OR THE STATEMENT OF THE PRINCIPLE.
USE UNAIDED		CLASSIFY OR CATEGORIZE OBJECTS, EVENTS, IDEAS, ACCORDING TO THEIR CHARACTERISTICS, WITH NO MEMORY AID.	APPLY THE STEPS OF THE PROCEDURE IN A SINGLE SITUATION OR ON A SINGLE PIECE OF EQUIPMENT, WITH NO MEMORY AID.	APPLY THE FORMULA OR RULE TO A VARIETY OF PROBLEMS OR SITUATIONS, WITH NO MEMORY AID.	USE THE PRINCIPLE TO EXPLAIN, PREDICT, OR DIAGNOSE WHY OR HOW THINGS HAPPENED OR WILL HAPPEN, WITH NO MEMORY AID.
USE AIDED		GIVEN CATEGORY CHARACTERISTICS AND GUIDELINES, CATEGORIZE OBJECTS, EVENTS, IDEAS, ACCORDING TO CHARACTERISTICS.	GIVEN STEPS OF THE PROCEDURE, APPLY THE PROCEDURE IN A SINGLE SITUATION, OR ON A SINGLE PIECE OF EQUIPMENT.	GIVEN THE FORMULA OR RULE STEPS, APPLY THE FORMULA OR RULE TO A VARIETY OF PROBLEMS OR SITUATIONS.	GIVEN A STATEMENT OF THE PRINCIPLE, EXPLAIN, PREDICT, OR DIAGNOSE WHY OR HOW THINGS HAPPENED OR WILL HAPPEN.

STEP 4: Determine whether or not the OBJECTIVE is APPROPRIATE?

4a: Are the CONDITIONS appropriate for the work to be performed on the job or for later training?

4b: Are the STANDARDS appropriate for the work to be performed on the job or for later training?

4c: Is the TASK LEVEL of the ACTION appropriate for the work to be performed on the job or for later craining?

4d: Is the CONTENT TYPE of the ACTION appropriate for the work to be performed on the job or for later training?

4e: If this objective is REMEMBER, is there a later USE objective?

4f: If this objective is USE-UNAIDED, is there a previous REMEMBER objective?

4g: If this objective is USE-AIDED, is the aid adequate, or are other objectives on the aid included?

Note, if the answer to 4d, 4e, or 4g is yes, and if the associated objective is be taught in the present course, evaluate that objective next and keep the related objectives together throughout the IQI evaluation.

STEP 2. Steps 2a and 2b refer to the CONDITIONS and STANDARDS parts of an objective. Several categories of conditions and standards are given. Obviously, no objective will require all of these. Each objective should be reviewed with these categories in mind, and a decision should be made about whether or not they are applicable. If you are not sure about whether or not a particular condition or standard should be included in an objective, a good rule is "when in doubt, stick it in."

As experienced instructional developers know, many objectives contain "implicit" conditions or standards, like "given paper and pencil" or "with 100% accuracy." It is up to each organization using the IQI to decide whether or not to include these obvious conditions and standards. Whatever the policy, though, it should be explicit.

Step 2c refers to the ACTION part of an objective. Obviously, the action part should use an action verb. It is usually best that there be only one action per objective; if there is more than one action, the objective should probably be split up into several objectives.

The action verb deserves special attention. It should always be an action that is observable and measurable. This means you should be able to tell whether the student did it or not. Action verbs like "appreciate" or "understand" are garbage words; who knows what they mean?

STEP 3. Refer back to the previous section for classifying objectives. If you can't classify an objective, the objective needs to be fixed.

If the objective fits in more than one box, it probably needs to be split up into more than one objective.

STEP 4. Steps 4a, 4b, 4c, and 4d all mean "REMEMBER THE JOB." The intent of any course is to prepare the student to do something after he finishes the course. This "something" is the JOB. The word JOB incorporates a wide range of activities, including on-the-job training, another more advanced course, or actual job performance. Therefore, the CONDITIONS, STANDARDS, TASK LEVEL, and CONTENT TYPE should be carefully evaluated, to make sure they are appropriate for the "job."

Steps 4e, 4f, and 4g are a partial check of the task analysis that led to the objectives. Step 4e means that there is no point teaching someone to remember something if it will never be used later. Note that "later" may be on the job or in a later course.

This is just as true for FACTS as it is for the other content types, but FACTS are not "used" in quite the same way. The information taught at the REMEMBER-level for CATEGORIES, PROCEDURES, RULES, and PRINCIPLES can be directly USED, but FACTS cannot. Instead, FACTS provide information to support all the other task/content types. Step 4e for FACTS therefore means "Is there some later objective that requires that the student know that fact information?"

Step 4f is the reverse of 4e. If an objective is USE-UNAIDED, this means that the student must remember what to do and then do it. Therefore, there should be a previous objective at the REMEMBER-level. Note that "previous" may be in an earlier course or may even be an entry behavior for the student. It should be emphasized that this step does not always need to be rigidly followed. In many Navy courses, there are laboratory exercises and/or performance tests that are in part based on previous REMEMBER-level objectives. That is, students are first taught the FACTS, and other REMEMBER-level information that they need to know to do the exercise or performance test. In many cases, the instructor or test administrator can determine whether or not the student has learned the REMEMBER-level information by how well he or she performs on the exercise or performance test. In this situation, it is usually unnecessary to test the REMEMBER-level information prior to testing the performance. It is usually more efficient to test both the REMEMBER and USE objectives at the same time. Care must be taken to construct the scoring keys for the performance test so that REMEMBER-level information is also evaluated. For example, for REMEMBER-PROCEDURE objectives, you can usually tell if a student remembers how to do a procedure just by watching his or her performance. Therefore, in most cases, a paper and pencil test of a REMEMBER-PROCEDURE objective is unnecessary. The only cases in which you might want to test memory for a procedure before observing performance is when the task is dangerous or involves expensive equipment, or when very precise diagnosis is required.

If the task level is REMEMBER, special care must be taken to make sure that the ACTION is appropriate. The reason for this is that there are really two kinds of remembering--recognition and recall. Recognition involves selecting or choosing from given alternatives, matching given pieces of information, or judging the accuracy of a given statement. In recognition, all the information is given; the student only has to make a decision about it. Recall, on the other hand, involves reproducing from memory some piece of information. Recognition and recall are different because recall involves more learning than recognition.

To make decisions about the appropriateness of recognition or recall, you must REMEMBER THE JOB. Most job situations require recall. For example, the steps of a USE-UNAIDED PROCEDURE must be recalled so they can be performed. The same is true for USE-UNAIDED CATEGORIES, RULES, and PRINCIPLES. Many FACTS also have to be recalled.

There are two situations in which recognition can be appropriate at the REMEMBER-level. The first occurs with FACTS, when a selection must be made from a group of objects, locations, etc. For example, the task "go to the tool box and get a ball-peen hammer" is a REMEMBER-FACT recognition task.

The second situation can occur for any content type at the REMEMBER-level when the job only requires the student to be generally familiar with the REMEMBER-level information. This situation only happens when the student is being prepared for later on-the-job or formal training and, even then, only when the student will be closely supervised. This is because the supervisor can take care of memory failure. For example, a student performing the steps of a maintenance procedure on a piece of equipment may not need to have memorized the steps if the supervisor was available to correct any errors or to explain what to do next. Even in this situation, it would have been more efficient and more consistent with the job if the student had been required to recall the steps of the procedure during training. If training time is limited, the recall performance criterion may be lowered when the job only requires "general familiarity."

The distinction between recognition and recall is important, because the type of testing to be done later and the type of instruction to be provided depends on whether the student is being trained to the recognition or the recall level.

Step 4g is a check on the quality of the job performance aid. There are a lot of terrible job performance aids in the world of work. Also, job performance aids may be hard to find or hard to use, or may use different technical vocabulary. In these cases, you may want to include objectives on how to use the aids or how to find them, or additional fact objectives on the vocabulary. In the worst case, the aid may be so bad that the objective must be rewritten as USE-UNAIDED. (Then you would also need a previous REMEMBER objective.)

Objectives Must Communicate

At the beginning of this chapter, we said that objectives must communicate to test developers, instructional developers, instructors, etc. This communication purpose must always be kept in mind when reviewing objectives, especially when subject matter expertise is not readily available. It is often desirable to include very specific, detailed descriptions of conditions, standards, and actions, so that later misunderstanding or errors do not occur.

For example, a complete objective may have to include not just a category, but a complete list of critical characteristics; all steps of a procedure or rule might have to be shown; a detailed description of a principle might be necessary. This could be accomplished by including references to documents that contain this information.

The reason why this might be necessaiy is that misunderstandings may occur if the instructional development team does not have sufficient subject matter expertise. For example, a test developer may have to see all steps of a procedure to write a good test item on it or to specify good scoring criteria.

Objectives Classification and Test Item Development

Again, objectives serve as the specification for test item development. For any objective, test items must be devised that consistently reflect the specifications in the objective.

First, the conditions of the test item or the conditions under which the item is administered must be developed to match the conditions stated in the objective. Naturally, there are some situations when, for reasons of safety, practicality, or cost, the testing conditions cannot be exactly the same as the conditions required in the objective. In these cases, it is important to simulate the conditions as closely as possible. In all cases it is important to REMEMBER THE JOB; that is, the testing situation must be close enough to the job situation, or later training situa.ion, to be sure that the student has achieved the objectives

Second, the standards in the test item or the standards for scoring the test item must be developed to match the standards specified in the objective. In criterion-referenced testing, standards are not arbitrarily selected. It makes no sense, for example, to require a student to get 80 percent of the items right if he needs to recall all the information to perform successfully on the job. On the other hand, for some tasks, a 70 or 80 percent criterion may be reasonable. In all cases, however, the standard specified in the objective should be used.

Third, the TASK/CONTENT level of the test item must match the TASK/CONTENT level of the objective. This usually means that the action verb in the test item should be the same as in the objective. At least, the same behavior must be required. If it is not, the test item is measuring something different than was required in the objective.

Implications of the Classification System for Test Item Development

The classification of an objective has further implications for the development of test items for that objective. One main distinction in the classification system is between REMEMBER-level and USE-level objectives. These two types of objectives have different testing requirements; thus, different item formats and different numbers of items will be required, depending on the task level and content type of the objective.

A similar distinction among types of objectives is made in NAVEDTRA 110A. NAVEDTRA 110A defines two types of objectives--terminal objectives and enabling objectives. Terminal objectives are specific statements of the performance expected from a student as the result of an experience, expressed in terms of the behavior to be exhibited, the condition(s) under which it is to be exhibited, and the standard to which it will be performed. Terminal objectives should translate directly to the tasks performed on-the-job. Enabling objectives are specific statements of the behavior to be exhibited, condition(s) under which it is to be exhibited, and the standard to which it will be performed. Enabling objectives must be written with conditions and standards appropriate to the training environment, and include knowledge/skills that support a terminal objective. USE-level objectives will almost always be terminal objectives and REMEMBER-level objectives will be enabling objectives. In addition, 110A defines learning steps. Learning steps are also almost always at the REMEMBER-level. The difference between learning steps and enabling objectives is that learning steps do not have to be tested while enabling objectives do. This handbook is only concerned with constructing test items and tests; therefore, learning steps will not be addressed. The decisions about whether REMEMBER-level information should be enabling objectives or learning steps should have been made during the analysis phase of the ISD process.

Item Formats

There are a number of different test item formats that may be more or less appropriate, depending on the TASK/CONTENT level of the objective. The following chart shows some acceptable formats for each level.

	CONTENT TYPE				
TASK LEVEL	FACT	CATEGORY	PROCEDURE	RULE	PRINCIPLE
REMEMBER	*for RECOGNITION* matching true-false multiple choice *for RECALL:* short answer fill-in listing	short answer fill-in listing	short answer fill-in listing	short answer fill-in listing	short answer fill-in listing
USE-UNAIDED		performance matching true-false multiple choice short answer fill-in	performance	performance true-false multiple choice short answer fill-in	performance true-false multiple choice short answer fill-in
USE-AIDED		performance matching true-false multiple choice short answer fill-in	performance	performance true-false multiple choice short answer fill-in	performance true-false multiple choice short answer fill-in

Item Formats at the REMEMBER-level

In the chart, notice that, at the REMEMBER-level, selected-response items (multiple-choice, matching, true-false) are usually not appropriate. This is because they don't test recall, only recognition. Most REMEMBER-level objectives require recall because of the nature of the job.

Earlier in this chapter, we discussed some situations in which recognition was appropriate for REMEMBER-level objectives. For these objectives, multiple-choice, matching, or true-false test items may be appropriate, even if the content type is CATEGORY, PROCEDURE, RULE, or PRINCIPLE. These do not appear on the chart, because even though they can be used, they are not the best choice. In this situation, it is also a good idea to recheck a recognition objective to make sure it is appropriate for the job. These issues are discussed in more detail in Chapter 3, which deals with REMEMBER-level test items.

Item Formats at the USE-level

Multiple-choice, matching, and true-false items can be appropriate, if carefully designed, for many USE-level tasks. For example, a category classification is often a true-false judgment. If the student must solve a math problem (USE-RULE), a multiple-choice item in which all alternatives are reasonable is appropriate. Also, some USE-PRINCIPLE predictions involve a limited set of possible alternatives; again, multiple-choice is appropriate.

Why is Format Important?

Test item format is important because students are not dumb! The first thing most new students do in a course is to find out how they will be tested. Then, they study just enough to pass the tests. If the objective requires a student to memorize something, multiple-choice tests should not be used, because students will learn just enough to recognize, not to recall. From your own experience, it should be clear that students study less carefully for a multiple-choice or true-false test than they do for a completion or short-answer test. The test items and the format should be like the tasks the student will do on the job.

Number of Items

Another distinction that can be made in the TASK/CONTENT classification system is between objectives that require the student to transfer knowledge to new situations and those that do not. (See page 22.) This distinction can be used to help determine the number of test items required by an objective.

At the REMEMBER-level, and at the USE-level for PROCEDURES, the student is required to recall specific pieces of information or to perform a specific procedural task the same way every time. Here, there is no requirement that the student transfer or generalize knowledge or performance to new situations. In these situations, there should ideally be at least one test item for each objective or, if one objective covers several pieces of information, there should be test items for all pieces. For example, if the student must remember the part names and functions of several pieces of equipment, the requirement could be incorporated in a single objective. But, there should then be enough test items to cover all the information specified. Similarly, if the objective requires a student to perform the steps of a procedure, then the test item(s) must assess performance of all steps of the procedure.

At the USE-level for CATEGORIES, RULES, and PRINCIPLES, the situation is different. These TASK/CONTENT levels mean that the student will be asked to apply knowledge to situations or problems that are new. Here, determining the required number of test items is more complicated. Since these objectives require transfer to new instances, there should be enough test items so that it is possible to decide if the student can apply what has been taught to new situations.

Chapter 4 of this handbook deals with test item development for USE-PROCEDURE objectives. Chapter 5 deals with CATEGORIES, RULES, and PRINCIPLES at the USE-level.

"Real-World" Constraints

Constraints on Testing

Although the prescriptions concerning consistency, format, and numbers of items described above should ideally always be followed, "real world" circumstances often result in insufficient budgets for test development, administration, and scoring. Although such constraints result in testing situations that are less than optimal, testing decisions still should not be made arbitrarily. The objective/job requirements and the constraints should be carefully analyzed, and test items should be designed so that the conditions, actions, and standards approximate as closely as possible the requirements of the objectives. Moreover, it is incumbent on the test developer to document what is being measured, what is not being measured, and what the effect is on the use or interpretation of test results.

Two types of constraints occur for tests of REMEMBER-level information: First, there is not enough time to test all the information required by the objective(s). Second, there are not enough resources for scoring, so that easy-to-score tests must be used.

For tests of USE-level performance, the same two constraints may apply. In addition, cost or safety may prohibit testing the real task.

These constraints, and how to deal with them, are dealt with in more detail in Chapters 3 through 5.

Constraints on Initial Specification of Objectives

Normally during ISD, objectives are developed on the basis of job/task analyses, and test items are developed later to be consistent with objectives. Sometimes, however, constraints on testing make it difficult or impossible to develop test items that are completely consistent with objectives. Unfortunately, some instructional developers then make the mistake of changing the objectives so as to maintain consistency with the test items. While the training and testing program may then be internally consistent, it is no longer consistent with the task analysis and the job.

This problem also occurs in a more subtle way during the initial specification of objectives. Sometimes, the developer has testing constraints in mind during objectives specification and lets these constraints affect the objectives. For example, suppose a developer knows that scoring constraints will require that multiple-choice tests be used. The developer may make the mistake of initially specifying objectives that require the student to "select from among four alternatives the correct definition of ...." rather than specifying objectives that require recall.

Objectives should specify the conditions, actions, and standards identified in the job analysis and the training analysis. Then, if testing constraints force deviation from the objectives, the deviations can be documented. This allows the instructional developers to justify requests for course modifications and additional resources. In addition, discrepancies between the original instructional intent and what is actually taught/learned and tested can be accurately reported to the operational community.

CHAPTER 3

TEST ITEMS FOR REMEMBER-LEVEL OBJECTIVES

Introduction

This chapter presents rules for designing test items for REMEMBER-level objectives. The first sections include discussions of when to test REMEMBER-level objectives and the differences between recognition and recall test items. Next rules for designing test items are given. The rules include methods for determining what test item formats should be and what type of information test items should test. After the rules are presented additional explanation for the rules and prototype items for each of the content types is provided. Next, rules and guidelines for writing the various types of REMEMBER-level test items are described. Finally, methods for assembling test items into tests and guidelines for dealing with constraints on testing are discussed.

Is It Always Necessary to Test REMEMBER-level Objectives Separately?

In Chapter 2 (page 29), we discussed situations in which testing REMEMBER-level objectives could be accomplished at the same time performance testing is done. This usually occurs when a laboratory exercise and/or performance test is included in the course. The exercise or performance test can often be constructed so that achievement of REMEMBER-level objectives can be evaluated at the same time. Therefore, before writing test items for the REMEMBER-level objectives, review the course plan to see if they can be tested during a laboratory exercise or performance test.

Recognition and Recall

As stated in Chapter 2, there are really two types of remembering; RECOGNITION and RECALL. Recognition involves selecting or choosing from given alternatives, matching given pieces of information, or judging the accuracy of a given statement. In recognition tasks, all the information is given; the student only has to make a decision about the information. Recall, on the other hand, involves reproducing from memory some piece of information. If an objective is well stated, it will be clear whether the objective requires the student to recognize or recall.

Just as there are two types of remembering, there are two main types of test items that can be used to test REMEMBER-level objectives. Test items for recognition objectives are called SELECTED-RESPONSE items; Selected-response items ask the student to choose the correct answer from two or more alternatives. The three most common varieties of selected-response items are TRUE-FALSE, MULTIPLE-CHOICE, and MATCHING items.

Test items for recall objectives are called CONSTRUCTED-RESPONSE items. A constructed-response item is one in which the student produces the answer, rather than selecting the answer from alternatives. The three most common types of constructed-response items are FILL-IN-THE-BLANK, SHORT-ANSWER, and LISTING items.

Selected-response Items for Recognition Objectives

Types of Selected-response Items.

1. *True-False Item*. A true-false item consists of a statement which the student must judge to be correct or incorrect, right or wrong, true or false, etc.

2. *Multiple-choice Item*. A multiple-choice item consists of (a) a stem that is the beginning statement, and (b) four or five or more response options. One of the options is the correct response. The other incorrect options are called "foils" or "distractors."

3. *Matching Item*. A matching item contains two columns, one consisting of stems, and the other consisting of response options.

Advantages of Selected-response Items.

1. *Easy to score*. Since student responses usually consist of a single letter or number, they can be quickly scored or scored by machine.

2. *Scoring is reliable*. Again, since responses are short and unambiguous, there is usually no disagreement about the correctness of a response.

3. *More questions*. More questions can be asked and answered in a typical testing period. Since students do not have to write out their answers, the time per item is reduced.

Disadvantages of Selected-response Items.

1. *Trivial*. Item writers tend to test trivial or irrelevant bits of information. (This disadvantage can be overcome if writers develop items for well-stated training objectives.)

2. *Difficult to develop*. Good items are difficult to develop. There are a number of errors that are often made in writing selected-response items. Some of these errors tend to make items easier to answer, and therefore less reliable as indicators of students' knowledge. Others make items confusing and also less reliable.

3. Easy to guess. Students have a significant chance of guessing the correct answer.

4. Few recognition. Few REMEMBER-level objectives in technical training are recognition objectives (most are recall). Even when a selected-response format is appropriate, true-false items are usually not appropriate.

Constructed-response Items for Recall Objectives

Types of Constructed-response Items.

1. Fill-in-the-blank Item. A fill-in item consists of a statement or question which requires a word or short phrase to be filled-in to complete the statement or answer the question.

2. Short-answer Item. A short-answer item consists of a question which can be answered with a few sentences or a single paragraph.

3. Listing Item. A listing item consists of a question which requires the student to give all the events, features, procedural steps, etc. necessary to answer the question.

Advantages of Constructed-response Items.

1. Easy to construct. A constructed-response item for a recall objective is usually easy to construct.

2. Difficult to guess. Students have little chance of guessing correct answers.

3. Most recall. Most REMEMBER-level objectives in technical training require recall. Recall can be tested most appropriately with constructed-response items.

Disadvantages of Constructed-response Items.

1. Difficult to score. Scoring is more time-consuming for constructed- response items. Although automated scoring is possible, it takes fairly complicated computer programs to do it.

2. Scoring is less reliable. Different scorers tend to score responses differently, unless they have been well trained. Scoring is subject to biases due to penmanship and students' writing ability.

3. Fewer questions. Fewer questions can be asked and answered, because students need time to write their answers.

Diagnostic Tests for REMEMBER-level Objectives

In Chapter 1, it was stated that diagnostic tests are used when it is necessary to determine in more detail the reasons for inadequate performance on a test item for a specific objective. For most REMEMBER-level objectives, however, it is usually not necessary to obtain additional diagnostic information. REMEMBER-level objectives require that a student recall a particular piece of information. If a student cannot answer a particular test item, this simply means that the student should study until the information can be recalled. Therefore, diagnostic tests for REMEMBER-level objectives are nearly always unnecessary. Further, such tests are also very difficult to construct.

Designing Test Items for REMEMBER-level Objectives

Rules for Designing Test Items

Rules for designing test items are given below.

Step 1. Check TASK LEVEL. Check the TASK LEVEL of the objective. If it is a REMEMBER-level objective, then continue; otherwise, refer to procedures for USE-level objectives.

Use the classification procedure given in Chapter 2 to determine the TASK LEVEL. This chapter deals with REMEMBER-level objectives. Procedures for designing test items for USE-level objectives are in Chapters 4 and 5.

Step 2. Determine CONTENT TYPE. Determine the CONTENT TYPE of the objective--FACT, CATEGORY, RULE, or PRINCIPLE.

Use the classification procedure given in Chapter 2 to determine the CONTENT TYPE of the objective. The CONTENT TYPE is necessary to use the charts in Steps 4 and 6.

Step 3. Determine RECALL or RECOGNITION Determine whether the REMEMBER-level objective requires RECALL or RECOGNITION.

Some guidelines for deciding whether an objective requires RECOGNITION or RECALL are:

a. Recall objectives will usually have action verbs like state, list, recall, name, etc. (See Chapter 2.)

b. Recognition objectives will have action verbs like recognize, match, choose, select, indicate, identify, etc.

c. In general, recall involves reproducing information from memory; recognition involves given information.

Step 4. Determine Item Type if RECALL Objective. If the objective requires RECALL, determine whether CONSTRUCTED-RESPONSE items can be used, or whether constraints on testing and scoring require that SELECTED-RESPONSE items must be used.

When it is necessary that tests be rapidly scored, or scored by machine, selected-response items may have to be used even though they are not entirely appropriate. When SELECTED-RESPONSE items must be used for recall objectives the following guidelines apply:

a. Do NOT change the objective. The objective must NOT be changed to reflect this constraint. As stated earlier, some designers make the mistake of changing the objective, rather than documenting the effect of the constraint on course effectiveness.

b. Document that an objective-test inconsistency will occur. It is important that deviations from appropriate test items be documented. This provides a basis for requesting additional testing resources so that tests can be made appropriate. In addition, discrepancies between the original instructional intent and what is actually tested can be accurately reported to operational commands and other users of test information.

c. See page 47 for design information for selected-response items.

When CONSTRUCTED-RESPONSE items can be used, use the guidelines below and chart on page 44 to specify type and content of items. Then, go to Step 7 on the following page.

a. The conditions under which the item is to be administered should be obtained from the objective.

b. The standards for scoring the test item should be obtained from the standards given in the objective.

c. The action for the test item should be obtained from the objective.

d. CONSTRUCTED-RESPONSE item should be paraphrased when possible. The rules for paraphrasing are given on page 53.

Step 5. Recheck Objective if RECOGNITION Required. If the objective requires RECOGNITION and the CONTENT TYPE is other than FACT, recheck the objective using the OBJECTIVE ADEQUACY procedure in Chapter 2.

> Note. As stated earlier, there are very few recognition objectives that are appropriate in technical training, and most of these are FACT objectives. If you have an objective that requires recognition of a CATEGORY, PROCEDURE, RULE, or PRINCIPLE, that objective should be rechecked for appropriateness using the objectives adequacy procedure in Chapter 2. Recall, instead of recognition, is usually more appropriate, unless mere familiarity with some topic is the goal of the instruction.

Step 6. Specify Design for RECOGNITION Objectives. If the objective requires RECOGNITION, or if a recall objective must be tested with selected-response items, use the chart on page 47 to specify type and content of items. Note that the chart requires the use of paraphrasing when possible.

Step 7. Design and Write Test Items. Use the guidelines, charts and procedures given in the following sections to specify the type and content of items and to write actual test items.

Specifying Type and Content of Items

Recall Objectives. The following chart should be used to specify type and content of constructed-response items.

TEST ITEMS FOR RECALL OBJECTIVES

FACT	CATEGORY	PROCEDURE	RULE	PRINCIPLE
1. Fill-in. 2. List.	1. List characteristics of category.	1. List steps of procedure.	1. List steps and decision. 2. Short answer. Give rule.	1. Short answer. Give principle.

Note. If several items of information are grouped in a single objective, then more than one test item may be required.

Prototype items for each of the content types are given below.

FACTS: Since Recall-FACT objectives typically require the student to recall names, dates, locations, part name, etc., the most appropriate item formats are:

Fill-in-the-blank: The student is asked to fill in a name, date, part name, etc.

"The (name, place, date, location, etc.) is ________"

Listing: The student is asked to give a series of names, dates, locations, etc.

"List the (names, dates, part-names, etc.) in the space below."

CATEGORIES: Recall-CATEGORY objectives typically require the student to recall definitions of category types. These definitions are lists of characteristics that define category membership. The recommended format is:

Listing: The student is asked to list category characteristics or to list category names and characteristics.

"In the space below, list the characteristics/features of an XXXX."

"In the space below, give the name and definition/features/ characteristics of each of the types of XXXX."

PROCEDURES: Recall-PROCEDURE objectives typically require the student to recall the steps of a procedure in order. As stated above, recall procedure objectives do not need to be tested if the student's memory for the procedure can be inferred from his performance of the procedural task. However, if the task is dangerous or involves expensive equipment or if precise diagnosis is required, memory for the procedure should be tested before the task is performed. The recommended format is:

Listing: The student is asked to list the steps.

"In the space below, list in order the steps for XXXX."

RULES: Recall-RULE objectives typically require the student to recall either (a) the steps to be followed in performing the rule-task, or (b) the general statement or formula that summarizes the rule. The recommended formats are:

Listing: The student is asked to list in order the steps to be performed and decisions to be made.

"In the space below, list the steps and decisions necessary to XXXX."

Short-answer: The student is asked to give a general statement or formula.

"In the space below, give the formula for calculating XXX."

"In the space below, summarize the rule/process for XXX."

PRINCIPLES: Recall-PRINICPLE objectives typically require the student to recall a principle statement; i.e., a statement of how or why something works the way it does. The format is:

> Short-answer: The student is asked to give a brief description of the operation of something, or reasons why something happened or it will happen, etc.
>
> > "In the space below, describe the operation of XXX."
> >
> > "In the space below give the causes and effects that lead to XXX."
> >
> > "In the space below, define the principle of XXX."

Recognition Objectives. The chart on the following page contains recommended formats for selected-response items for recognition objectives. Note that the chart requires paraphrasing when possible.

TEST ITEMS FOR RECOGNITION OBJECTIVES

FACTS:

1. Multiple-choice. Items should be paraphrased for events but not for part names, proper names, dates, etc.

2. Matching.

CATEGORY:

1. Multiple-choice. There should be one item for category characteristic

2. Matching. Items should be written as follows. "Column A contains a list of categories and Column B contains a list of characteristics. Match each category with its characteristics."

PROCEDURES:

1. Multiple-choice. Items should be paraphrased if possible. There should be one item for each step and the distractors should be other steps or common errors.

2. Matching. Matching can be used if there are more than five steps and if the test is not machine-scored.

RULES:

1. Multiple-choice. This type of item can be used to test for recognition of the general rule or formula. It can also be written like multiple-choice items for procedures to test for recognition of each step. If the latter type of items is to be used the items should be paraphrased if possible.

2. Matching. Similar to matching for procedures.

PRINCIPLES:

1. Multiple-choice. Paraphrasing should be used.

Prototype items and explanatio:ı for each content type are given below.

FACTS: Recognize-FACT objectives usually require the student to associate some given name, date, part-name, etc. with some other given information. Paraphrasing should be used except for dates, proper names, or technical vocabulary. Recommended formats are as follows:

Multiple-choice: The student is asked to choose the alternative associated with something given in the stem.

"The (event) happened in

a. (date) b. (date) c. (date)"

"The symbol for the XXX is

a. (symool) b. (symbol) c. (symbol)"

Matching: The student is asked to associate each item given in Column A with something in Column B.

"Column A below gives part-names. Column B contains the functions of parts. Match each name with its function."

"Column A below gives major events. Column B contains dates. Match each event with its date of occurrence."

CATEGORY: Recognize-CATEGORY objectives require the student to recognize characteristics that define category types or category membership. Paraphrasing should not be used for category names but can be used for category characteristics unless technical vocabulary must be used. Recommended formats are as follows:

Multiple-choice (single-answer): The student is given a category name in the stem, and a list of characteristics, one of which is a correct characteristic.

"A critical characteristic of a (category name) is:

a. (irrelevant characteristic)
b. (correct characteristic)
c. (irrelevant characteristic)
etc."

Multiple-choice (multiple-answer): The student is given a category name in the stem, and a list of characteristics, several of which may be correct. The instructions are to choose all that apply. This item format requires at least one item for each category being tested. Note that this type of item requires special scoring for both hand and machine scoring.

> "Which of the following are essential characteristics of a (category name)? (Choose all that apply.)
>
> a. (correct characteristic)
> b. (correct characteristic)
> c. (irrelevant characteristic)
> etc."

Matching (single-answer): If several categories are being tested, and if each category has a single definition/characteristic/feature, then a standard matching item can be used. Column A contains category names. Column A contains category definitions. The student is asked to match each name in Column A with the correct definition in Column B.

> "Column A contains names of (category types). Column B contains definitions/characteristics/features. Match each category in Column A with the correct definition in Column B.

A	B
1. Category name 1	A. Definition for 2.
2. Category name 2	B. Definition for 1.
etc.	etc."

Matching (multiple answer): If several categories are being tested, and if the categories have multiple characteristics some of which are common to more than one category, then a "multiple matching" item can be used. Column A contains category names; Column B contains correct and irrelevant characteristics. The student is asked to match each name in Column A with as many characteristics in Column B as apply. Note that this type of item requires special scoring for both hand and machine scoring.

> "Column A contains names of (category types). Column B contains characteristics/features. For each item in Column A, choose as many items in Column B as apply. Answers in B may be used as many times as necessary.

A	B
1. Category name 1	A. Characteristic X
2. Category name 2	B. Characteristic Y
3. Category name 3	C. Characteristic Z
etc.	etc."

PROCEDURE: Recognize-PROCEDURE objectives require the student to recognize the steps of a procedure and recognize the order in which the steps should occur. Paraphrase techniques should be used for all steps except for technical vocabulary. Recommended item formats are:

Multiple-choice: In the stem, the student is instructed to choose the alternative that is a correct statement of the first, second, third, etc., step in the procedure. The distractors should contain incorrect statements of steps, and steps that are correct but that are not the first, second, etc., step. This format requires one item for each step in the procedure.

> "Which of the following is a correct statement of the second step in the process for XXX?
>
> a. (correct statement of Step 3)
> b. (correct statement of Step 2)
> c. (incorrect statement of Step 2)
> etc."

Matching: Column A contains step numbers. Column B contains descriptions of procedural steps, and includes incorrect statements of steps. The student is asked to match each step number in Column A with the correct description of that step in Column B. This format requires an item for the procedure but should not be used with procedures that are extremely lengthy.

> "Match each step number in Column A with the correct statement of the step in Column B.
>
> Procedure for XXXXX.

Column A	Column B
1. Step 1	a. (correct description of Step 3)
2. Step 2	b. (incorrect description of Step 4)
3. Step 3	c. (correct description of Step 4)
	etc."

RULES: Recognize-RULE objectives require the student to recognize the steps and decisions in each rule, and they require the student to recognize correct descriptions or statements of the formula or rule summary. Paraphrase should be used for all steps and decisions except for symbols or technical vocabulary.

If rule steps and their order are being tested, then the item formats are the same as procedures.

If recognition of a formula or summary statement is required, then a multiple-choice item that requires the student to choose the correct formula or statement from distractors should be used.

If decision steps in rules are being tested, these should be treated as category tasks (or facts if they are simple). The reason for this is that decision steps are really classification decisions. In these cases, if the decisions are not simple, they should be analyzed as a category task, and the item formats are the same as for categories.

PRINCIPLE: Recognize-PRINCIPLE objectives require the student to recognize cause-and-effect or predictive relationships or explanations of these relationships. Paraphrase should be used except for technical vocabulary. The recommended item format is:

> Multiple-choice: The student is asked to determine which of a given series of statements best explains the principle of relationship.
>
> > "Which of the following statements best describes the principle of XXXX?
> >
> > a. (correct description or explanation)
> > b. (incorrect description)
> > etc."

Rules and Guidelines for Writing Test Items

Paraphrase Test Items

What is Paraphrasing? Paraphrasing means to change the words in a test item without changing the meaning.

Why should Paraphrasing be Used? Paraphrasing is used to make sure that the student has learned the meaning of what has been presented and not simply memorized words. According to Anderson (1972):

> "in order to answer a question based on paraphrase, a person has to have comprehended the original (information), since a paraphrase is related to the original (information) with respect to meaning but unrelated with respect to the shape or the sound of the words."

When Should Paraphrasing be Used? Paraphrasing should be used whenever possible. Of course, there are situations in which the student must recall verbatim information, such as technical vocabulary, names, dates, or exact steps of procedures or statements of rules or principles. In all other situations, however, it is the meaning that should be tested. It is particularly important to use paraphrase in selected-response items, because recognition of words that have been seen before is easier than recognizing meaning.

Rules for Paraphrasing. Several techniques for creating paraphrases have been given by Anderson (1972):

Step 1: Change all substantive words (nouns, verbs, modifiers) to other words with the same meanings. It is not necessary to change only single words and vice versa.

Step 2: If possible, rearrange the phrases within a sentence or the sentences within a series of sentences. That is, change the words or phrases around so that they are not in the same order but so that the meaning is the same.

EXAMPLES OF PARAPHRASING:

Example 1:

Original test item:

"Which of the following is a recommended technique for creating paraphrase test items?

A. Change all substantive words (nouns, verbs, modifiers) to other words with the same meanings.

B. Choose synonyms for all technical vocabulary terms.

C. Spell certain words incorrectly but so that they sound the same."

Note that the original test item and the correct alternative (A) were taken verbatim from the instruction. To construct the paraphrase, the content words in the stem and the correct alternative were changed; e.g., "test items" was changed to "exam questions." Also, the sentences were reordered.

Paraphrased test item:

"To develop exam questions that test for meaning, it is best to

A. Keep the information the same, but use different terminology (subject and action words, adverbs, adjectives).

B. Choose synonyms for all technical vocabulary terms.

C. Spell certain words incorrectly but so that they sound the same."

Note that in multiple-choice items, it is not necessary to change the incorrect alternatives, unless they are also taken verbatim from the instruction. In this case, the alternatives were made up, and did not come from the instruction.

Example 2:

Original test item:

"The BASE of TRANSISTOR Q2 connects to the___________."

In this case, no real paraphrase is possible, because the item consists mainly of technical terms.

Guidelines for Writing True-False Items

The primary guideline for true-false items is that they are not appropriate and should not be used. However, if it is absolutely necessary to use true-false items, follow these guidelines.

1. Only a single idea or piece of information should be tested in a true-false item. If a test item contains more than one idea or piece of information, and if one is true and the other false, the student may not know how to answer the question. Items should be either completely false or completely true. Also, if a student fails an item, it is difficult to determine which piece of information the student does not know.

2. Avoid shades of meaning. Items should deal with information which is clearly true or false, no opinions or "judgment calls."

3. A true-false item should be a simple positive statement; negatives should be avoided. Negatives make items harder to read so they take longer to answer. Negatives also make items less reliable; they can be missed because of confusion and not because the students do not know the content. Consider the following sentence: "Avoiding negatives does not make the item harder to read."

4. Don't use words like "never," "always," "sometimes," etc. Items which include words like "never" or "always" are usually false because there are exceptions to most rules. Items which include words like "sometimes" or "rarely" or "usually" are usually true because these items are usually exceptions.

5. Items should be kept short. Long items are harder to read, and therefore more difficult to judge true or false. Also, long items often suffer from other problems described above.

EXAMPLES OF TRUE-FALSE ITEMS:

Example 1:

Original test item:

"A lieutenant does not wear either two gold sleeve stripes or a single gold-bar collar device." (T or F)

Note that this item contains the negative "not" and also tests more than one idea. This makes the item quite confusing. The revised item tests only a single idea, and uses positive language.

Revised test item:

"A Navy lieutenant wears two gold sleeve stripes on each sleeve of the winter blue uniform." (T or F)

Example 2:

Original test item:

"Admiral Halsey's decision to abandon Leyte Gulf during World War II to pursue a carrier force to the north was probably justified by circumstances." (T or F)

Note that this item deals with shades of meaning. Although Halsey thought the decision was justified, other historians do not agree. The revised item asks for the information as stated in a reference book.

Revised test item:

"According to your naval history textbook, Halsey should have abandoned Leyte Gulf." (T or F)

Guidelines for Writing Multiple-choice Items

Guidelines for writing multiple-choice items are listed below.

1. Wording in the stem should be clear and unambiguous, so only one answer is correct.

2. If a negative is used in the stem or alternatives, it should be highlighted for emphasis. Negatives should generally not be used for the same reasons as for true-false items. If negatives must be used, however, they should be highlighted so that students notice them and interpret the item correctly.

3. All repetitive phrases are included in the stem, rather than being stated over and over in the alternatives.

4. All alternatives should be grammatically consistent with the stem, have a similar grammatic structure, and be approximately the same length. Alternatives that are not grammatically consistent with the stem are usually incorrect and can therefore be eliminated. In addition, when alternatives are not similar in grammatical structure and length, learners may lean toward or away from the correct answer. For example, a longer answer is often correct because it has information not included in the other alternatives.

5. Each alternative should be a believable completion of the stem. If an alternative is not believable nobody is likely to choose it.

6. Alternatives should not be paraphrases or synonyms of one another and alternatives that are opposites should be used with caution. A paraphrase or synonym for the correct answer would also be a correct answer, so paraphrases should be avoided. Opposites are frequently easier to come up with than other alternatives, but they are often inappropriate, because they are not believable alternatives.

7. Alternatives such as "all of the above," "a and b above," and similar alternatives should not be used. It is definitely easier to put one of these in as a final alternative than to think of something better, but they are limited in their usefulness because they can often be easily eliminated. "None of the above" should not be used if the student is asked to select the "best answer" rather than the correct answer.

EXAMPLES OF MULTIPLE-CHOICE ITEMS:

Example 1:

Original test item:

"A typical malfunctioning oil purifier will emit some:

A. blue smoke.
B. whining sound.
C. periodic 'thunk.'
D. all of the above."

Note that the alternatives in this item are not grammatically parallel. No one will choose "emit some whining sound" as a correct alternative. Also, "all of the above" should not be used.

Revised test item:

"A typical malfunctioning oil purifier will emit:

A. blue smoke.
B. a whining sound.
C. a periodic 'thunk.'
D. colorless gas."

Example 2:

Original test item:

"Lubricating oil for the steam propulsion plant aboard ship

A. should be changed frequently.
B. should be changed every 2000 miles.
C. should be changed when the plant starts to emit a periodic 'thunk.'
D. should be changed every 30 days."

Note that the alternatives in this item are not all believable and also contain repetitive phrases. No one is likely to choose "when the plant starts to emit a periodic thunk." In addition, the phrase "should be changed" should be included in the stem.

Revised test item:

"Lubricating oil for the steam propulsion plant aboard ship should be changed

A. every 500 miles.
B. every 1000 miles.
C. every 1500 miles.
D. every 2000 miles."

Guidelines for Writing Matching Items

Guidelines for writing matching items are as follows.

1. Directions for matching items should include:

    a. An explanation of what is in each column.

    b. A statement of how entries in the two columns should be matched.

    c. How often choices in Column B may be used.

    d. How many answers are possible for each entry in Column A.

    Here is a set of directives for a typical matching item. Notice how each of the points above are stated.

    > "Match the electronic schematic symbols in Column A with the electronic components for which they stand in Column B. There is only one answer for each symbol in Column A, and no component in Column B can be used more than once."

2. There should be extra entries in Column B (unless choices can be used more than once).

3. The entries in each column should be similar. If you can't put a brief statement or title defining each column, then you probably have more than one type of entry. There should be separate items for different types of things.

4. The entries in each column should be concise and clearly worded.

EXAMPLES OF MATCHING ITEMS:

Example 1:

Original test item:

"Match the columns below:

1. International ship call sign
2. International shore call sign
3. Task Organization call sign
4. Voice call sign
5. Publication where call signs are listed
6. Security classification of Voice call signs

A. "N" + 4 letters
B. "N" + 3 letters
C. "N" + 2 letters
D. letter-number-letter-letter
E. Confidential
F. ACP-126
G. word (or words)
H. Secret."

Note that the directions for this item are insufficient. They don't specify what is in the columns or how matching is to be done. Furthermore, the entries in the columns are not similar. The manual where call signs are located is not related to the definitions of call signs.

Revised test item:

This item has improved directions and similar entries in each column.

"Column A below contains types of call signs. Column B below contains definitions for different types of call signs. Match each type in Column A with its definition in Column B. Answers in Column B may be used only once, and there is only one answer for each type in Column A.

A. Types of Call Signs

1. International ship call sign
2. International shore call sign
3. Indefinite call sign
4. Task Organization call sign
5. Voice call sign

B. Definitions

a. "N" + 1 letter (except "R")
b. "N + 2 letters
c. "N" + 3 letters
d. "N" + 4 letters
e. letter-number-letter-letter
f. word (or words)
g. "N"-number-letter"

Guidelines for Writing Fill-In-The-Blank Items

Guidelines for writing fill-in-the-blank items are as follows.

1. Only the things that are important for the learner to remember are left blank. Trivial and unimportant words are not left out. The word that is left out determines what the item is measuring. If trivial words are omitted, trivial things are being measured.

2. The item is worded so that only one word or phrase correctly completes the sentence. Items for which almost any answer would make sense are avoided. If the item is worded so that more than one answer makes it a true complete sentence, any such answer can be justified as correct. It is important, therefore, to word items so that learners must supply the intended correct answer to complete the sentence accurately.

For example, the item "Columbus discovered America in ____" could be correctly completed with "the Santa Maria." If the item was intended to measure the year of discovery, it should say so. A better wording would be "Columbus discovered America in the year ____."

3. Grammatical cues or other cues to the correct answer should be avoided. Grammatical cues can help a student who does not know the correct answer to get the item right. For example, in the item "A fill-in item should have at most ____blank," the only reasonable answer is "one" because the work "blank" is singular.

4. The correct answer not be a "giveaway" word which could be guessed by someone who does not really understand the information. For example, in the item "Long items take more time to answer than ____items," the first word that comes to mind is "short."

5. The blank should be placed near the end of the item. Items with the blank near he beginning are harder to read and take longer to answer.

6. There should be only one blank in a single item. More than one blank often results in an item with little meaning. For example, in the item "The symbol for ____is ____," any pair containing a symbol and what it represents is correct.

7. In the scoring key, all acceptable synonyms or alternative correct answers should be specified. Acceptable answers should be determined by subject matter experts. Answers given by students during the tryout of the test items can also be used to identify acceptable correct answers.

EXAMPLES OF FILL-IN-THE-BLANK ITEMS:

Example 1:

Original test item:

"In a ____ item, the blank should be placed near the_____ of the item."

Note that this item has more than one blank. Also, the first blank is near the beginning. Finally, the word "blank" in the item gives away the first answer, because no other item types have blanks.

Revised test item:

"In a fill-in item, the blank should be placed near the ___ of the item."

Example 2:

Original test item:

"It is best not to use selected-____ items for Remember-level objectives."

Note that this item tests trivial information. The important point is tested in the revised item.

Revised test item:

"Selected response items should not be used for ____ -level objectives."

Guidelines for Writing Short-answer Items

Guidelines for writing short-answer items are as follows.

1. The question should be clear and complete, and should provide enough information so that the student who knows the answer can produce a correct one.

The learner should know what is expected. The question should tell the learner clearly what information should be included in the short answer.

2. The required answers should be short; that is, no longer than one or two sentences or phrases.

3. The directions to the student should indicate how the items will be scored without giving cues to the correct answer. The learner should be provided with exact information concerning how performance will be graded. This should include such things as number of points for each part of the answer, whether partial credit will be given for partial answers, or whether a complete answer must be given.

4. The scoring key for the test grader should include all acceptable synonyms or alternatives or forms of the correct answer. Acceptable answers can be determine by subject-matter experts. Response given by students during the tryout of the test items can also be used to aid in identifying acceptable answers.

5. If there is only one possible answer for a short answer item, it might be more appropriate to change it to a fill-in-the-blank item. Short—answer items are most appropriate when learners have some flexibility in giving the correct answer.

EXAMPLES OF SHORT-ANSWER ITEMS:

Example 1:

Original test item:

"Describe paraphrasing in the space below."

Note that the original item is not clear and complete. Does the item ask how to paraphrase, to describe when paraphrasing is important, or what? Also, the original item contains no scoring information.

Revised test item:

"Describe Anderson's (1972) method for constructing paraphrases of fill-in-the-blank test items. To receive credit, your answer must include all the steps that Anderson recommends."

Example 2:

Original test item:

"In the space below, discuss th rules for constructing fill-in test items."

Note that the original item is is not clear. How is the student to know what "discuss" means? Also, the item does not include scoring information.

Revised test item:

"In the space below, discuss the rules for constructing fill-in test items. Your answer should include each rule given in your text and should also include a brief explanation of why each rule is important. You will receive one point for each each correct explanation."

Guidelines for Writing Listing Test Items

1. The question should tell the student what the contents of the list should be. Students should know exactly what is expected of them. They should be told if they are supposed to list a portion of a larger list. In addition, they should be told if they should give additional information about the items in the list.

2. If a specific number of things must be listed, the question should specify this number (unless the number would be a hint).

3. The directions should indicate how the item will be scored without giving hints to the correct answers. Students should know if they must list items in a certain order, if they must use the exact words given in the instruction, and/or if some parts of the list are more important than others. They should not be given information that will help them generate the list from the directions alone.

4. If sequence is important, the scoring key should judge sequences separately from completeness. Partial credit should be allowed. If one item is out of order, this could result in no credit for an answer that is otherwise correct and complete. Instead, sequence should be scored separately, and students should be told this in the directions.

5. The scoring key should specify allowable alternatives if any and should allow for different scoring weights for more or less important items on the list.

EXAMPLES OF LISTING ITEMS:

Example 1:

Original test item:

"In the space below, list all the test items you have learned."

Note that the original item does not clearly state what the student is to list, or how many there are. No scoring information is given.

Revised test item:

"In the space below, list the six types of test items given in the chapter on remember-level test items. You will be given one point for each correct type."

Example 2:

Original test item:

"List the blocks of the ISD process."

Note that the original item does not state how many blocks must be listed, and does not state whether order is important. Scoring information is not given.

Revised test item:

"List IN ORDER the 5 blocks of the ISD process. You will be given one point for each correct block, and an additional five points if they are all in the correct order."

Assembling Tests for REMEMBER-level Test Items

Building Tests

Once test items have been written for all REMEMBER-level objectives, the items must be assembled into module tests, lesson tests, etc. This is a straightforward process; simply group objectives/test items together in related topics according to the lessons, modules, etc. of the course. If there is some logical order in which objectives are taught (for example, if one topic is prerequisite to another), then these topics should be tested in the same order. If no logical ordering is required, items can be arranged in any order.

In many cases, alternate forms of tests are necessary to minimize cheating or simple memorization of answer keys. Alternate forms can be developed by randomly ordering the test items, unless test item order is important. Another technique for selected-response items (multiple-choice, matching) is to rearrange the alternative responses. A third technique for developing alternate test forms is to paraphrase items using the paraphrase rules given earlier in this chapter.

The next step in building tests is to develop scoring keys. Because the goal of testing is achievement of the learning objectives, the scoring keys should reflect the standards specified in the objectives. Remediation prescriptions should be built into the scoring key, so that students who do not achieve the objectives can be remediated efficiently. This is done simply by keeping track of which items test which objectives, and where the instruction or remedial material for those objectives is located. In most cases for REMEMBER-level tests, it is more efficient to set a standard for an entire test or for groups of related objectives within a test. Remediation in this situation is only given to students who do not achieve the test standard(s). A procedure for setting test standard(s) for REMEMBER-level tests is given in the following section. This procedure ensures that all essential objectives are tested, and remediated and retested if necessary.

Finally, instructions for the test should be developed. The instructions should give students general information about how to answer the types of items contained in the test (e.g., multiple-choice, short-answer, etc.). The instructions should also list any technical manuals or course documents that are the basis for the correct answers.

Setting Standards

Navy training is concerned with the achievement of objectives. Therefore, test standards will normally be those that are specified in the objectives. The standards specified in the objectives should be those required by the job. This is because the logic of ISD requires that job performance requirements be reflected in training objectives and test items. This section is concerned with setting standards based on job requirements for entire tests or sections of tests. It is important to realize that setting standards for REMEMBER-level objectives and tests is not as straightforward as setting standards for USE-level objectives and tests. This is because, at the REMEMBER-level, we are dealing with enabling objectives that support the job but that are not "performed" on the job. Therefore, setting standards is more difficult and more arbitrary.

Although there are several techniques available for setting test standards, they all have at least two things in common. First, the procedure for setting the standard(s) must be carried out prior to actually using the testing instruments. This is done because, with criterion-referenced testing, the standard(s) must be independent of the students' performance on the test. There is a strong temptation, for example, to set a fairly low standard if many of the first group of students get fairly low scores on a test covering a newly-developed piece of material. Standard(s) must be based on a consideration of the relative importance of the subject matter and not on consideration of student performance on the test. They must, therefore, be set before the tests are used.

Second, setting test standards for REMEMBER-level tests is somewhat arbitrary. There exists no "scientifically objective" or completely valid way to set test standards for REMEMBER-level (enabling) objectives and test items. But "arbitrary" is not necessarily "unacceptable." Many other very important standards in life are set arbitrarily, on the basis of judgment. If reasoned judgment by qualified people is used as a basis for a somewhat arbitrary standard, the standard will be valid for separating students into appropriate groups. On the other hand, if randomness or convention is used as a basis for a standard, the standard will cause poor decisions to be made about remediation and retesting.

The procedure recommended in this manual for developing test standards for REMEMBER-level tests is a modification of one which was designed by Robert L. Ebel (1979). It is one of the more objective procedures available and not so complex that sophisticated computers and software are needed. This approach uses the judgment of a group of experts in the subject matter area (five such judges are recommended). It is based on rating test items for job relevance and item difficulty. Test items are rated instead of objectives because a given objective may have several test items associated with it. Thus, rating individual test items results in a more precise standard.

Procedural Steps.

Step 1. The most critical aspect of the procedure is that the raters clearly understand, in specific terms, the use to which the standard(s) will be put. The standard is used to separate students into two groups--those to be retested and those for which retesting is not required. The raters who set the standard must be told this.

Step 2. The raters must know what is expected of a student who will not be retested. For example, if the knowledge information is extensive and will not be completely covered on a follow-on test, then the student who will not be retested should probably be required to score very well on the test. On the other hand, if the knowledge information will be covered and scored on a later performance test, then the standard could be set at a lower level. In all cases, the situation must be explained in detail to the raters before they are presented with the test items.

Step 3. Each judge is required to determine, for each item, how relevant it is to the job the student will be expected to do, as well as its level of difficulty. This is done by assigning each item to one of the six categories shown below.

RELEVANCE/DIFFICULTY MATRIX

		Difficulty Levels		
		Easy	Medium	Hard
Relevance Categories	Essential		--a	--a
	Important			--a
	Acceptable			

a-Level of difficulty is not a factor for consideration with essential items. For important items, only two levels of difficulty should be specified.

The following definitions of the three categories of item relevance should be provided to the raters:

Essential. Students must know this material. They cannot and should not be allowed to pass the test if this item is not answered correctly. Raters must be in complete agreement on essential items.

Important. This material covers very important aspects of the job or is important prerequisite knowledge needed for further training.

Acceptable. This material is a valid part of what will be expected of the passing student, however, the student does not have to master it completely to perform the job or go on to further training.

The meaning of "easy, medium, hard" is difficult to define precisely. Level of difficulty usually has to do with the complexity of the material or amount of knowledge or skill required for the student to select the correct response. Additionally, the way the test item is written may increase the level of difficulty. Item length, choice of words, use of grammar, etc. all can make a test question more or less difficult. Some definition of complexity and an explanation of how items can be written to be more or less difficult should be given to the raters who categorize the items. It should be noted that item difficulty can be confirmed after the exam has been given a few times. The P statistic discussed in Chapter 6 is a measure of difficulty.

Step 4. Once each of the judges has placed each of the items in one of the six categories in the relevance/difficulty matrix, the judgmental part of the procedure is completed. The number of times an item is placed in a category is then multiplied by a weighting factor developed by Ebel. These six products are added and the sum is divided by the total number of placements of all items by all judges (e.g., 5 judges placing 50 items each would produce a denominator of 250). The quotient oecomes the standard for the test.

The six weighting factors for the Relevance/Difficulty Matrix are as follows:

Matrix Categories	Expected Success
Essential	100%
Important	
Easy	90%
Medium	70%
Acceptable	
Easy	80%
Medium	60%
Hard	40%

An example of how this procedure would work for a 50 item test using 5 judges is represented in the following table:

	No. of item Placements/ by 5 Raters	Expected Success (percent)	Placements X Exp. Success = Cut Score
Essential	50	100	5000
Important			
Easy	56	90	5040
Medium	77	70	5390
Acceptable			
Easy	15	80	1200
Medium	25	60	1500
Hard	26	40	1040
TOTAL	250		19170

STANDARD = 19170/250 = 76.68 OR 77%

It is important to note that this procedure allows students to answer some essential items incorrectly and still score above the standard. Therefore, the scoring and remediation instructions must be written so that any essential items that are missed are retested.

A second event to watch out for is if an item or items receive low ratings by all raters. These items may cover contextual or background information that is not really necessary to the performance of the terminal objective. If this is true, the objectives associated with these items should be changed from objectives to learning steps.

There are many other techniques for setting standards. Most are more complex than the procedure described above. For a more sophisticated approach that includes a look-up table of values, see Brennan (1981).

Constraints on Testing at the REMEMBER-level

There are two main types of constraints on REMEMBER-level testing that affect both test item development and building tests from items:

1. Due to lack of resources for test administration and scoring, quick easy-to-score tests must be used.

2. There is not enough time to test all the information required by the objectives.

The first constraint was dealt with earlier in this chapter, where we recommended the use of paraphrased selected-response items in cases where constructed-response items were required but couldn't be used.

The second major constraint on testing at the REMEMBER-level is that not everything can be tested because there is insufficient time. The obvious strategy for dealing with lack of time is to sample from the information to be learned. There are two different approaches to sampling: sampling-by-topics and sampling-by-importance. Sampling-by-topics is used when all the information to be learned is equally important. Sampling-by-importance is used when some information is more important for later job performance than other information. These approaches are described below.

Sampling by Topics. When all the information is important to learn, pieces of information must be sampled randomly within topics. The reasons we divide subject matter into topics, instead of just choosing items randomly from the entire content, are because:

1. It is important that all the information be "covered."

2. It is important to identify particular topic areas in which the student may need remediation.

Dividing the information to be tested into topics make testing and remediation more efficient and accurate. When a student misses an item, we don't have to remediate and retest everything, just the information from the topic area for the missed item. It should be obvious, therefore, that many "narrow" topics are better than a few broad ones.

Information can be grouped into topics by asking a subject matter expert to organize the content into "logical" areas. For example, if a student must remember 100 different names and locations of controls on a piece of equipment, these might be grouped into "alignment and adjustment" controls, "power supply" controls, "test-mode" controls, etc.

When the content has been divided into topic areas, test items can be sampled from each area. The number of items sampled from each area for the test should reflect the total number of items in that area; that is, topic areas with a large number of total items should have more items selected for testing than topic areas with a small number of total items. A good rule to follow when selecting items for the test is that the percent of the total number of items covered by each topic area should be the percent of the total number of test items allowed for that topic area. This sampling process is called "stratified random sampling."

If sampling is used in a test because of time constraints, this has important implications for the instruction that teaches to that test. In a normal program, where everything is tested, the test items "cover" everything that is taught. Students, therefore, must study all the material to achieve the objectives. In a sampling situation, however, not all the information is tested. Conceivably, students could study only some of the material and still pass the test. To ensure that students do study everything, the most effective strategy is to develop several forms of the test that, together, cover all the objectives. Students are then told that they may be tested on any material, and their best study strategy is to learn ALL the material. Further, practice questions during instruction should be designed to lead the students to study the type of information to be tested, not just the specific information on any one test.

The standards for tests that are constructed by sampling should still be the standards specified in the individual objectives. If everything could be tested, we would require students to achieve all the objectives and sampling doesn't change this.

Sampling by Importance. Many training programs include REMEMBER-level information, not all of which is equally important to test. In these cases, when time constraints prevent all information from being tested, testing should be concentrated on the most important information. Less important information should be tested less extensively.

The first step in sampling by importance is to determine which information should be tested extensively and which should not. The way to do this is to ask a subject matter expert to review the objectives and establish a priority for each. Such criteria as safety, possibility of damage to expensive equipment, effect on success of a mission, and "need to know" versus "nice to know" should be examined.

After the objectives/test items have been prioritized, tests can be assembled simply by sampling many of the high priority items and only a few of the lower priority items. The tests should include as many of the high priority items as possible, and low priority items should be included as time permits. Alternate forms of the tests should use different samples of low priority items.

Standards for high priority items should require students to achieve each high priority objective. Standards for other items should reflect the priority determined by subject matter experts. Low priority items could have a standard lower than those specified in the objectives because, by definition, they are not critical.

This sampling-by-importance scheme also has implications for instruction and remediation. Since we want students to achieve each objective for high priority information, students should be told which information is important and what the standard will be. Students should also be told about the standard for less important information, so they can spend their study time appropriately. Remediation should be carefully planned for high priority items so that students are retested and can achieve the objectives. Less extensive remediation without retesting can be used for lower priority information.

CHAPTER 4

TEST ITEMS FOR USE-PROCEDURE OBJECTIVES

Introduction

As stated in Chapter 2, USE-PROCEDURE objectives typically require students to perform a specific procedural task the same way every time. Here, there is no requirement that students transfer or generalize their knowledge or performance to new situations. In other words, everything the student needs to know or do can be taught and tested. Other tasks that do require transfer or generalization are dealt with in Chapter 5.

Testing USE-PROCEDURE Objectives

Several factors must be considered in designing test items for USE-PROCEDURE objectives. These include:

1. Whether the product that results when the procedure is performed, or the process used to accomplish the procedure, or both, must be tested.

2. How much diagnosis is required.

3. Whether memory for the procedural steps is to be measured by observing performance instead of by a paper and pencil test.

4. Whether a simulation should be used to test the procedure.

5. What test item format and scoring method (checklist or rating scale) should be used.

The following pages contain rules for designing and developing test items for USE-PROCEDURE objectives.

Designing Test Items for USE-PROCEDURE Objectives

Rules for Designing Test Items

The following rules are used for USE-PROCEDURE objectives regardless of whether the task is AIDED or UNAIDED. The only difference will be whether or not the aid is provided during testing.

Step 1. Determine TASK LEVEL and CONTENT TYPE. Check the TASK LEVEL and CONTENT TYPE of the objective. If it is USE-PROCEDURE (AIDED or UNAIDED), then continue; otherwise, refer to the chapter appropriate for the task level and content type.

Step 2. Determine if the Objective Specifies a Product. Determine whether or not the objective specifies a product or accomplishment. A product is a perceptible result (something you can see, hear, smell, or touch) of a procedure. A product may or may not be specified in the objective. Product measurement is possible when:

1. The objective specifies a product.

2. The product can be measured as to presence or absence or according to its characteristics.

3. The steps of the procedure can be performed in different orders without affecting the product.

Examples:

1. The student will set up and connect a Simpson 260-5P multimeter for measuring resistance. (Note that here, one can look at the multimeter after the procedure is performed to determine if it is correctly connected and set up.)

2. The student will correctly complete OPM Form SF-71, "Application for Leave."

Step 3. Determine Whether to Test Product, Process or Both. If the procedure does result in a product, then it is necessary to determine whether to test the product, the process or both. Guidelines for making this decision are given below. If, after using these guidelines you decided to measure process, go to step 5 for rules for analyzing processes. If you decide to measure the product, go to step 6 for rules for analyzing products.

1. If the objective contains specific standards the product must meet, the product should be evaluated. Similarly, if the objective gives specific standards that must be followed during the process, the process should be evaluated. Standards might include safety procedures, time standards, requirements that the procedural steps be performed in a certain order, etc.

2. If it is not necessary to measure both process and product, then select the easiest to measure. Factors that should be considered in making this selection are:

a. The time it takes to do the measurement.

b. How many personnel are required to do the measurement.

c. Whether the product can be accurately measured without looking at the process.

d. Whether errors made early in the process could be costly or dangerous.

3. If diagnosis is important (i. e., if it is important to know exactly when and where the errors occurred so that remediation can be given), the process should be measured. Also, if memory for the procedural steps is to be measured by observing performance, the process should usually be measured. Product measurement may be appropriate only if it is possible to detect when and where any possible error occurred just by measuring the product.

4. If the procedure does not result in a product, then the process must be measured. Process measurement can be done when the objective specifies a sequence of performances that can be observed or when the behavior itself is the result in which you are interested.

Process measurement is appropriate when:

a. The product and the process are the same thing (as in giving a speech).

b. There is a product, but safety, high cost, or other constraints prevent the product from being measured.

c. It is necessary to diagnose reasons for performance failure.

d. There may be a product, but there are critical points in the action sequence which must be performed correctly because of the possibility of damage to personnel or equipment.

Examples: Process measurement can be used on the following objectives:

1. Given a globe valve, rags, prussian blue, gasket material, packing, and tools, the student will correctly disassemble and reassemble the valve. (Note that since the globe valve would have to be disassembled again to determine if the student correctly reassembled it, it is more efficient to measure the student's process.)

2. Given a knife, a match, a shoestring, and a scarf, the student will demonstrate first aid for a snakebite when the victim develops severe symptoms. (Note that although a product--a dead or live victim--results, it is the process of administering first aid that is important.)

Step 4. Determine Whether Simulation is Necessary or Desirable. There are some situations in which a simulation should be used to test performance of a procedure. A simulation occurs when at least one of the conditions, actions, or standards required for on-the-job performance is changed in the testing situation. It should be noted that because USE-PROCEDURE tasks consist of a simple series of steps, it is usually a simple matter to design simulations for them. (The situation is different for CATEGORY, RULE, or PRINCIPLE). However, if an objective requires product evaluation rather than process evaluation, simulaticn cannot be used because a simulated procedure does not generate the same product that the real-world procedure would. There are two reasons for using simulation to test performance of a procedure.

1. There may be constraints on the testing situation which make it impossible to test the procedure as it is performed on the job. These constraints include (a) lack of equipment, (b) lack of personnel to monitor or participate in testing, (c) insufficient time, (d) safety of personnel, or (e) risk of damage to equipment.

2. It may be desirable to simulate even if the procedure could be tested as it is performed on the job. Normally a requirement for simulation during testing that occurs because of the constraints listed above will be specified in the objective. Often, however, situations in which it may be desirable to simulate will be overlooked during objectives development. For this reason, objectives should be reviewed to determine if simulation is desirable, and if so, they should be revised. Reasons why it may be desirable to simulate include the following.

a. It may be possible to save time, equipment, or personnel.

b. Time can be spent on critical steps. Simulations can often be accomplished in less than "real time." Unimportant steps or equipment start-up time can be skipped.

c. Simulators can be designed to record more performance/ diagnostic data than can be obtained from real equipment. The simulation can also be designed to be "played back" so that the student can critique his own performance.

d. Test situations can be standardized.

Examples: The examples below give objectives that do not specifically call for a simulation but for which simulation may be desirable.

1. Given a knife, a match, a shoestring and a scarf, the student will demonstrate first aid for a snakebite when the victim develops severe symptoms. (Note that it would probably be unwise to have an actual victim with severe snakebite symptoms as part of the test situation. Furthermore, the first aid procedure involves making incisions across the bite area. This test item can be simulated using a plastic "victim," and/or by using a felt-tip pen to indicate where the "incision" should be made.)

2. Given a Caterpillar bulldozer, the student will demonstrate start-up and shut-down procedures, and procedures for raising and lowering the bucket. (Note that a bulldozer is expensive to operate, and is usually not available for training purposes. This could be simulated using a mock-up of the controls.)

Step 5. Analyze the Process. The following guidelines and rules should be used to analyze the process into steps.

1. List in sequence the steps a skilled person completes when performing the process.

2. It may be necessary to break down the steps into "sub-steps" depending on:

a. The complexity of the steps.

b. The possibility of damage to equipment or personnel.

c. The types of diagnostic information desired.

d. Whether or not memory for the steps is to be measured by observing performance.

e. The ability of the people to be tested.

3. Refine the steps by observing other experts and reviewing the steps with them. It should be emphasized that this analysis is critical for developing test items and that it cannot be accomplished without the help of a highly qualified subject matter expert.

Step 6. Analyze the Product. The following guidelines and rules should be used to analyze the product to determine what qualities should be measured.

1. Collect many samples of the product and sort them into good and bad categories.

2. Go through all the good samples and note the characteristics they have in common that make them good.

3. Go through all the bad samples and note the characteristics that make them bad.

4. Refine the good and bad characteristics in consultation with job experts. Again, the analysis is critical and cannot be accomplished without the help of a subject matter expert.

Step 7. Determine Whether to Use Checklists and/or Rating Scales. There are two primary ways to measure procedures: by using checklists and by using rating scales. Both can be used to measure either processes or products. A checklist and/or rating scale should be developed for each step of the process or each characteristic of the product. Guidelines for deciding whether to use checklists or rating scales for both processes and products are given below.

1. For each step of a process, use the guidelines below to determine whether a rating scale or a checklist is the appropriate measurement method.

a. If the step is either done or not done and performance can be measured by simply checking "yes" or "no," a checklist should be used.

b. If performance of a step can vary in quality from high to low, best to worst, good to bad, or some other scale, a rating scale should be used. For example, olympic judges use rating scales for diving and gymnastic events.

c. Rating scales can also be used when a step has more than two possible outcomes.

d. It is not necessary that all steps use the same measurement method; some may use checklists and others may use rating scales.

2. For each characteristic of a product, use the guidelines below to determine whether a checklist or rating scale is the appropriate measurement method.

a. If the component or quality characteristic of the product is either present or absent and can be measured by simply checking "yes" or "no," a checklist should be used.

b. If the product can vary in quality from high to low, adequate to inadequate, good to bad, or some other scale, a rating scale should be used.

c. It is not necessary that all the characteristics of the product use the same measurement method, some may use checklists and others may use rating scales.

Step 8. Develop Checklists. The following guidelines should be used to develop checklists for products and processes.

1. Checklists for process measurement are used to determine whether or not a step of the procedure has been performed. To construct a checklist for process measurement, use the behavioral descriptions of the procedural steps obtained in the analysis in Step 5. Constructing a checklist from these descriptions is simple. Each description of a step is checked to determine whether or not the behavior has been performed. For some tasks, certain steps may not need to be performed each time. For these steps, the checklist should include an "not applicable" or N/A check as well as the yes and no checks.

Example: The following is a checklist for the process of brewing a pot of coffee:

	yes	no
1. disconnect coffee pot		
2. disassemble coffee pot		
3. clean pot and components		
4. inspect components		
5. fill pot with water		
6. reassemble components		
7. fill basket with coffee		
8. reconnect coffee pot		
9. set dial on coffee pot		
10. report pot is perking properly		

2. Checklists for product measurement are used to determine whether a characteristic or aspect of the product is present or absent. To construct a checklist for product measurement, use the descriptions of the characteristics obtained in the analysis in Step 6. Each description of a characteristic is checked to determine whether the characteristic is present or absent. Characteristics that should be absent are stated negatively.

Example: The following is part of a checklist for evaluating a typed letter:

yes no

1. correct top margin....................................
2. correct ieft margin...................................
3. correct right margin..................................
4. date properly positioned at top-right.................
5. salutaticn has comma or colon.........................
6. no lines too long or short............................

Step 9. Develop Rating Scales. The following guidelines should be used to develop rating scales for process and product measurement. The first three guidelines are general guidelines for rating scale development. Guidelines four and five address process and product measurement respectively. This step concludes with some general comments about both checklists and rating scales. It is important to read these comments.

1. Rating scales are used when it is necessary to make finer judgments about a process step or a product characteristic that can be obtained from a checklist. Rating scales are used to make judgments about how well, quickly, accurately, etc. a process step has been performed, or about the degree to which a product characteristic varies in quality, tolerance, accuracy, etc.

2. Rating scales differ from checklists in that rating scales contain more than just two response options. The number of response options depends on the process or product to be measured, but there should not usually be more than about nine options. It is also important that a rating scale only rate one action. If several actions are included on one scale, the scale is confusing and difficult to use.

Example: The following is an example of a rating scale that includes too many actions and is therefore not a good rating scale.

RATING SCALE FOR SWIMMING ABILITY

Very poor form and speed. Thrashes in water. Unable to complete one length of pool.	Below average form and speed. Major flaws in arm action. Head too high or low in water. Breathes inefficiently. Irregular path.	Average form and speed. Breathes from one side only. Several flaws in arm or kick action.	Form and speed are good. Minor flaws in arm and kick action and head position.	Form and speed are excellent. Follows line. Looks ahead. Water forehead level. Breathes right and left. Excellent arm action.
1	2	3	4	5

Each of the actions in this scale should have been broken out in separate scales. For example:

RATING SCALE FOR BREATHING DURING SWIMMING

Breathes inefficiently with no regular pattern of head movement.	Breathes regularly from one side only.	Breathes regularly alternating right and left.
1	2	3

3. In addition to containing only one action rating scales must have a description of observable behaviors for each response option. It is very important that a rating scale contain such a description for each response option, because these descriptions make the rating scale easier to complete and much more reliable. You will occasionally see rating scales that do not contain descriptions for each option, like the following:

RATING SCALE FOR SWIMMING ABILITY

poor					good
1	2	3	4	5	6

Scales like this should never be used, because it is not clear how they should be used. Different judges will give different ratings for the same thing because there are no descriptions to "anchor" the ratings. Anchoring is important because the purpose of rating scales is to diagnose what students can and cannot do. If the descriptions are not detailed enough, it will be difficult to prescribe appropriate remediation.

4. Rating scales for process measurement are used to determine the degree to which a step of a procedure has been performed well. To construct a rating scale for process measurement, use the behavioral descriptions of the procedural steps obtained in the analysis in Step 5. The number of response options, and the "anchoring" descriptions for each option, should be obtained from standards given in the procedure or objective, or from a subject matter expert. Each rating scale should rate only one action.

Example: Suppose one step in a first-aid procedure is to perform mouth-to-mouth artificial respiration. Rating scales like the following may be constructed:

Give Artificial Respiration at the Proper Rate.

Unable to perform artificial respiration.	Rate is too slow (less than 10/minute.	Rate is once every five seconds. (12/minute)	Rate is too fast (over 15/minute.)
1	2	3	4

Inflate Victim's Chest

Fails to clear airway or hold nose closed. No air reaches victim's lungs.	Victim's chest is overinflated, and excessive pressure is blown into victim.	Victim's chest is not inflated enough.	Fills victim's lungs so chest is inflated enough.
1	2	3	4

5. Rating scales for product measurement are used to determine the degree to which a product characteristic meets requirements. To construct a rating scale for product measurement, use the descriptions of good and bad characteristics obtained in Step 6. The number of response options and the "anchoring" descriptions for each option are obtained from standards in the objective or procedure or from a subject matter expert. The guidelines for product measurement are similar to those for process measurement. A rating scale should only rate one characteristic.

Example: Suppose the product to be measured is a cabinet that has been constructed and painted, and that two of the characteristics to be measured concern the quality of the paint job. Two rating scales like the following might be constructed.

Quality of Paint Finish

Paint film is too thin, and is not uniform.	Paint film is too thin, but is uniform.	Paint film is too thick.	Paint film is uniform, has correct thickness
1	2	3	4

There are light spots which are unfinished.	Buffing is good, but color is too pale.	Finish doesn't shine.	Finish contains runs, sags, or drips which have been buffed over.	Finish is buffed to mirror shine.
1	2	3	4	5

A FINAL WORD ABOUT CHECKLISTS AND RATING SCALES: The decision about whether to use a checklist or a rating scale is often arbitrary; that is, it often does not make a difference whether a checklist or rating scale is used. The reason for this is that almost all rating scales can be turned into checklists, and some checklists can be made into rating scales. The important thinq to remember is to define the checklist steps and the rating scale decisions as precisely as possible. The more completely required behaviors are described, the more effective the checklist or rating scale will be. It is also important to note that the numbers used with rating scales are only a convenience for checking reliability. They are not used to score students. The anchoring statements are used to score students. They are important for scoring because they are used to diagnose what the student can and cannot do. This diagnosis is accomplished by carefully defining the response options in the anchoring statement of the rating scale.

Step 10. Develop Directions for Administering the Rating Scales and/or Checklists. Good directions for using checklists and rating scales are particularly important, because checklists and rating scales rely on the user's judgment. While this subjectivity can make ratings unreliable or error-prone, good directions can reduce this by controlling administration and scoring and by specifying controlled conditions under which the measurement is used.

Directions should be developed to specify exactly how the rating scales and/or checklists are to be used and scored. Here are some sample directions:

> You are being asked to evaluate trainee performance using the rating scale/checklist below. Rate or check off the trainee's performance according to the descriptions of the behaviors (product characteristics) provided on the scale or checklist. Indicate your judgment by placing an 'X' in the box for the description that most accurately reflects your evaluation.

Instructors who will use the checklists or rating scales should at least be trained by making them familiar with the directions and giving them sufficient explanations of the characteristics of the processes or products to be evaluated. Furthermore, if the rating scales or checklists call for difficult discriminations or fine distinctions among characteristics, instructors should be given an opportunity to practice evaluating actual processes or products and should be given feedback about how accurately they perform.

Step 11. Field Test and Revise the Rating Scales and/or Checklists. Procedures for ensuring that rating scales and checklists are reliable across raters are described in Chapter 6.

Step 12. Determine How Many Repetitions of the Procedure are Required. At this point, you need to decide how many times the student must perform the procedure during testing. Use the following guidelines to make this decision:

1. If the procedure is complicated or involves intricate motor movements, it should be tested more than once. A subject matter expert should be consulted to determine the appropriate number of times the test should be given.

2. If there is a time limit in which the procedure must be accomplished, the procedure should be tested more than once. Again, a subject matter expert should be consulted.

Step 13. Develop REMEMBER-level Test Items if Necessary. If REMEMBER-level objectives will be tested during the USE-PROCEDURE test, develop the necessary test items using the procedures described in Chapter 3.

REMEMBER-level test items may be given orally or on a written test and may be administered during or prior to the USE-PROCEDURE test. REMEMBER-level test items should not be given during the USE-PROCEDURE test if they interfere with the performance of the procedure. This could occur if the procedure had to be accomplished in a given amount of time or if several steps had to be performed one right after the other.

REMEMBER-level items given during a USE-PROCEDURE test should test memory for part names, part locations, part purposes and functions, symbol names, procedural steps, and definitions of unfamiliar words and technical terms.

EXAMPLE: Designing Test Items for a USE-PROCEDURE Objective

The example described in this section illustrates the entire process for developing test items for a USE-PROCEDURE objective. The objective is as follows:

> Objective. Given assorted tools, rags, a cotter pin, prussian blue, gasket material, packing, and a 2-inch low-pressure globe valve, the student will disassemble, inspect, and reassemble the globe valve, and perform an operational test, in accordance with the procedure given in the rate training manual. The student will observe safety precautions during the process.

The steps for designing a test item for the above objective are given below.

Step 1. Determine TASK LEVEL and CONTENT TYPE. As stated, this objective is a USE-UNAIDED PROCEDURE.

Step 2. Determine if the Objective Specifies a Product. The objective calls for an assembled globe valve. This could be considered a product; however, it may be difficult to evaluate without observing the process. A subject matter expert should be consulted.

Step 3. Determine Whether to Test Product, Process, or Both. This objective does not specify explicit standards the product must meet, other than the implicit standard that the globe valve will "pass" the operational test. There are, however, standards specified concerning the process: It must be performed in accordance with the MRC procedure, and safety precautions must be observed. In addition, in this case, the operational test is part of the process the student is supposed to perform; it is not a product check that the instructor does. Therefore, only process measurement is required for this objective.

Step 4. Determine Whether Simulation is Necessary or Desirable. For this procedure, simulation is unnecessary and impractical. Globe valves are fairly cheap and portable. There is little risk of safety or damage because these are low-pressure valves, and this procedure is typically performed on valves that have been removed from their system. Further, there appears to be no easy way to simulate this task that would reduce the need for instructors or the time to perform the task.

Step 5. Analyze the Process. The first thing to do is to analyze the procedure into steps. The steps of this procedure are listed below:

(1) Disassemble the valve.

(a) Remove handwheel.
(b) Remove packing gland nut and bushing.
(c) Remove bonnet.
(d) Remove stem from bonnet.
(e) Remove packing from stuffing box.

(2) Inspect the seat, disc, and stem for cuts, burrs, or scratches.

(3) Conduct a spotting-in test of the disc and seat using prussian blue.

(4) Reassemble the valve in reverse order of disassembly.

(5) Repack the valve.

(6) Perform operational test.

In this case, some of the steps above are broken down enough, while others need more refinement. For example, Step (3) above (spotting-in) needs more explanation. These steps are given below:

(3) Conduct a spotting-in test.......

(a) Coat seating area of disc with thin film of prussian blue.
(b) Reassemble (but do not repack) valve.
(c) Close valve so disc seats as it would in normal operation.
(d) Disassemble valve.
(e) Inspect seat to determine if prussian blue has been transferred to seat uniformly around 100% of disc seating area and across at least 85% of seating area width.

Finally, all steps are refined in consultation with a subject matter expert. In this case, it was found that the steps to be performed when the seat, disc, or stem fail inspection in step (2), or when the spotting-in test in step (3) is negative are not included. The missing steps that should be added are:

(2.1) If the seat or disc has cuts, burrs, or scratches, lap the valve as in Step (3.1). If the stem has cuts, burrs, or scratches, machine the valve stem.

(3.1) If the spotting-in test fails, then lap the valve using lapping compound.

(a) coat disc with lapping compound.
(b) reassemble (but do not repack) valve.
(c) twirl the stem so as to grind the disc against the valve seat.
(d) disassemble valve.
(e) conduct another spotting-in test.

Step 6. Analyze the Product. This step can be skipped because the process is being measured.

Step 7. Determine Whether to Use Checklists and/or Rating Scales. Next, after the steps of the procedure are sufficiently refined, it is necessary to determine whether a checklist or rating scale should be used for each step.

In this example, substeps (1a) through (1e) (disassemble the valve) can be measured with a checklist, because these steps are simply performed or not.

Step (2) (inspect seat, disc and stem) requires a rating scale, because the step can be performed correctly, performed incorrectly in two different ways, or not performed at all.

Step (3) (conduct the spotting-in test) requires rating scales for substeps (3a) and (3e), and a checklist for substeps (3b), (3c), and (3d). For substep (3a), "Coat seating area of disc with thin film of prussian blue," it is important that the right amount of prussian blue be applied. If too much is used, cuts, burrs, or scratches will not be detected. If too little is used, the test may fail unnecessarily. Substep (3e) "Inspect seat ..." requires a similar rating scale to ensure that the disc and seat were inspected accurately.

Steps (4), (5), and (6) require checklists because each of the steps is either performed or not.

Steps (2.1) and (3.1) above both require checklists.

Steps 8 and 9. Develop Checklists and Develop Rating Scales. The next step is to develop checklists and rating scales for each of the process steps identified in step 5 above. The checklist for substeps (1a) through (1e) is shown below.

	performed	not performed
1. Disassemble the valve.		
a. Remove handwheel....................................		
b. Remove packing gland nut and bushing....................................		
c. Remove bonnet....................................		
d. Remove stem from bonnet....................................		
e. Remove packing from stuffing box....................................		

The rating scale for Step (2) is:

2. Inspect the seat, disc, and stem for cuts, burrs, or scratches.

Did not inspect.	Failed to identify cuts, burrs, or scratches.	Identified cuts, burrs, or scratches that were not present.	Performed inspection accurately.

The rating scale for substep (3a) is:

3. Coat seating area of disc with thin film of prussian blue.

Did not apply prussian blue.	Applied too little prussian blue.	Applied too much prussian blue.	Applied correct amount of prussian blue.

The remaining checklists or rating scales for the rest of the steps are developed similarly.

Step 10. Develo Directions for Administering the Rating Scales and/or Chec' ' 'ts. Next, directions for administering and scoring the checklists/rating scales are developed. Directions for the current example are as follows:

> You will evaluate student performance in disassembling, inspecting, reassembling, and testing the 2-inch low-pressure globe valve using the checklists and rating scales provided. For the checklists, simply check whether or not the step was performed. For the rating scales, check the box that most accurately describes the student's behavior. To pass, the student must perform all steps correctly. If the student fails to perform a checklist step, refer him or her back to the training manual. If the student performs a rating scale step incorrectly, you may give more instruction yourself and retest the student later.

If the instructors for this objective are subject matter experts, they may be able to use the checklists and rating scales without training. If the instructors are not expert, they will need training. For example, the instructor will need to be able to tell if the correct amount of prussian blue is used in substep (3a).

Step 11. Field Test and Revise the Rating Scales and/or Checklists. This step requires field tests and revision of the checklists or rating scales. This process is described in chapter 6.

Step 12. Determine How Many Repetitions of the Procedure are Required. For this example, after consulting with subject matter experts, it was determined that the procedure was not complicated and there were no time standards. Therefore, the task only had to be done once.

Step 13. Develop REMEMBER-level Test Items if Necessary. In this example, no REMEMBER-level objectives will be tested, so no REMEMBER-level test items need to be developed.

CHAPTER 5

TEST ITEMS FOR USE-CATEGORY, USE-RULE, AND USE-PRINCIPLE OBJECTIVES

Introduction

As stated in Chapter 2, CATEGORY, RULE, and PRINCIPLE objectives at the USE-level require the student to deal with new cases not seen during training. That is, the student is required to generalize or transfer knowledge or performance to new situations.

USE-CATEGORY objectives require the student to classify, sort, or identify a large number of possible objects, events, etc. into one of a small number of particular categories. Instead of having to remember each object and its classification, the student is taught to use the characteristics that define a category to classify items not seen before.

USE-RULE objectives require that a large number of problems be solved or a complicated series of steps be performed on a large number of different objects, events, etc. Instead of having to remember each problem or the steps for each different object, the student is taught a rule that can be used to deal with problems not seen before.

USE-PRINCIPLE objectives require explanation, prediction or diagnosis of a large number of possible situations, events, effects, etc. Instead of having to remember each possible situation or event and its effects, the student is taught a principle that explains "how" or "why" situations or events occur. The student can use the principle to explain or predict or diagnose a variety of situations not seen before.

The key words "not seen before" in the paragraphs above indicate the requirement for transfer or generalization of training.

Testing for Transfer is Different

Developing tests for objectives that require transfer is different than developing tests for objectives that do not. In the previous chapters for REMEMBER and USE-PROCEDURE objectives, each objective "defines" only one or a small number of possible test items, and these items are "seen before" during training. For example, if an objective requires the student to recall the names of the parts of a piece of equipment, the only way to test this is to ask the student to recall the names of the parts, and the student will have practiced recalling these names during training. Similarly, if an objective requires the student to perform a procedure on a piece of equipment, the test will require the student to perform the procedure, and the student will have practiced this performance during training.

In contrast, a USE-CATEGORY, USE-RULE, or USE-PRINCIPLE objective "defines" a large (or even infinite) number of test items. Only some of these can be practiced during training, and only some can be given on tests. For example, if an objective requires the student to use a rule that applies to a large or infinite number of problems, some of these will be used as examples or practice items during instruction, and others will appear on tests. There will still be many problems "left over" that could have been tested and that the student is supposed to be able to solve.

The major difference between testing objectives that require transfer and testing objectives that do not is that transfer objectives involve a large number of possible test items, not all of which can actually be tested. This situation results in the following problems that must be addressed when developing tests for objectives that require transfer:

1. What is the "domain" or set of test items "defined" by the objective?

As stated above, REMEMBER and USE-PROCEDURE objectives define only one or a small number of test items, while USE-CATEGORY, USE-RULE, and USE-PRINCIPLE objectives define large or infinite numbers of test items. This large set of test items is called the item "domain" for the objective. (Tests for these objectives are sometimes called "domain-referenced" tests.) The problem here is to analyze the objective to determine what the domain of items is.

2. Once the domain is determined, how should items be chosen from the domain?

Since not all items from the domain can be administered, which ones should be? Items should be chosen so that a "fair" or representative sample of the domain is obtained. If the sample accurately represents the domain, then a student's score on the sample should accurately represent the score the student would get on a test containing all the items in the domain if such a test were possible. In other words, if a student gets all the items in the sample correct, can you be sure that he would get any new items from the domain correct? This depends on how representative the sample of items is.

3. If the student does not achieve the standard specified in the learning objective on the sample of items, how can weakness be diagnosed and remediation prescribed?

Again, since not all items from the domain can be administered, how can you tell from a student's performance on a few items what the student doesn't know or can't do? If the domain is analyzed appropriately, it is possible to pinpoint areas of student misunderstanding with just a few test items. Then, remediation can be given efficiently.

4. Once the test is constructed, how should standards be established?

Standards for "domain-referenced" tests are set the same way standards are set for other types of criterion-referenced tests. Standards are established on the basis of the job requirements and are specified in the objectives. (See Chapter 1, "Setting Standards for Criterion-referenced Tests.")

It is important to realize that a good analysis of the domain or subject matter is the basis for answering these questions. In the following sections, we will describe some techniques for analyzing the CATEGORY, RULE, and PRINCIPLE content types. We will also discuss appropriate item formats for the three content types. Finally, we will present methods for dealing with constraints on the analysis, item formats, and the availability of equipment necessary for testing. For all three content types, the same general plan will be used to develop test items:

Step 1. Analyze the Objective to Determine the Test Item Domain.

Step 2. Determine Item Formats.

Step 3. Construct Actual Test Items.

Step 4. Develop Standards and Instructions for Scoring and Diagnosis.

Step 5. Develop REMEMBER-level Test Items if Necessary.

The next three sections deal with these issues for CATEGORIES, RULES, and PRINCIPLES respectively.

Testing USE-CATEGORY Objectives

As stated earlier, all USE-CATEGORY objectives involve presenting the student with instances or objects to be sorted, identified, classified, or categorized according to type. Each type is defined by critical characteristics or features. If an object has the right combination of features or characteristics, it is an EXAMPLE or member of the type defined by those features; otherwise it is a NONEXAMPLE. To test a USE-CATEGORY objective adequately, it is necessary to analyze the category specified in the objective to determine what the critical characteristics are and then to construct items that test all those characteristics.

Categories are defined by critical characteristics, but they also have characteristics or features that are NOT critical. These are called "variable" characteristics. They are frequently associated or correlated with critical characteristics but cannot be used to define a category type, because they cannot always be used to separate examples from nonexamples. Students who make mistakes in identifying examples and nonexamples of a category are often looking at the variable characteristics instead of the critical ones. For example, it is true that most dogs have hair, but so do many other animals, and some dogs do not have hair. Therefore, "having hair" is not a critical characteristic of the category "dog"; instead, it is a variable characteristic. A student who uses the variable characteristic "hair" to classify animals as if "hair" were a critical characteristic will make errors. When developing tests for USE-CATEGORY objectives, it is important to identify all the correlated variable characteristics, and include them in the test items.

Both critical characteristics and variable characteristics must be identified so that tests can be DIAGNOSTIC. If the test items are selected carefully, the scorer will be able to tell from test item responses whether a student is attending to all critical characteristics, or is being misled by variable characteristics. The steps needed to analyze categories and develop test items and diagnostic tests are described on the following pages.

Step 1. Analyze the Objective to Determine the Test Item Domain

Step 1A. List the Critical Characteristics that Determine Category Membership. If a good task analysis has been done, the critical characteristics will have been identified, and there will be an enabling "remember" objective for them. (There may even be enabling objectives for the characteristics if they are unfamiliar or require difficult discriminations.) Otherwise, it will be necessary to analyze the category task to determine the characteristics. Consult a subject matter expert.

For example, suppose we have an objective that requires students to categorize pieces of furniture (or pictures) according to type (e.g., tables, chairs, sofas, etc.). Let's consider one of these categories, the category "table." The critical characteristics could be (1) it has a flat surface, (2) the surface is horizontal, and (3) the surface is supported from the floor. These characteristics are determined either by obtaining some standard definition of pieces of furniture or by looking at lots of pieces of furniture and attempting to determine what characteristics define tables as opposed to other pieces of furniture. That is, one lists all the features or characteristics that tables must have but that other pieces of furniture may or may not have.

Step 1B. List Exceptions (if any). There may be exceptions for some categories. For example, an object may be included in a category "by definition" even though it lacks a critical characteristic, or some object may be "arbitrarily" excluded from a category even though it has all critical characteristics. These exceptions should be rare. If a category has lots of "exceptions," it may be better to redefine the category and change the objective.

In our "table" example, there are a few exceptions. For example, a few work tables, like drafting tables, may not have a horizontal top; instead, they are sometimes tilted. Most tables, however, do have horizontal tops. These tilted work tables are exceptions; hopefully, there are not too many of them. If there are a lot of exceptions, our critical characteristics for tables would have to be redefined.

Step 1C. List the Variable Characteristics and the Values These May Take. Both examples and nonexamples of a category will have characteristics or features or properties that can vary without affecting category membership. These characteristics are "variable." Variable characteristics are usually NOT identified during a task analysis, so it will be necessary to determine them. One way to do this is to collect lots of examples of the category, and list the characteristics they do NOT have in common.

To identify the variable characteristics in our "table" example, we would look at lots of tables to determine what features they have that can vary without changing the table to something else. Among the variable characteristics of tables are (1) number of legs--tables may have one pedestal leg, two legs, four, or even more, (2) the material the table is made of--wood, plastic, metal, stone, etc., (3) the actual height--coffee tables are low, kitchen tables are medium, work tables are higher, etc., and (4) the shape of the table--some are round, some are square, some are kidney-shaped, etc.

It is important to determine the variable characteristics because they can cause students to make errors. Therefore, a test should use items that contain variable characteristics that typically lead to confusion or common errors. That is, errors that are commonly found on the job or in the world. This means that a "common error" analysis should be conducted. This is done by analyzing the mistakes that job performers typically make to determine what characteristics of the task are responsible for the errors. These characteristics then become the variable characteristics in the test items.

Step 1D. Use the Critical and Variable Characteristics to Determine the Diagnostic Requirements and Specify the Types of Items Needed. For category tasks, a student can perform accurately or inaccurately for a variety of reasons. The test must help determine whether the student is performing accurately and, if not, why. That is, the test must be diagnostic.

A diagnostic test for a category task must sample all the ways in which a student could respond. When a student is classifying objects, there are four situations that can occur:

1. The student classifies a "true example" correctly as an example of the category.

2. The student classifies a "nonexample" correctly as a non-example.

3. The student classifies a "true example" incorrectly as a non-example.

4. The student classifies a "nonexample" incorrectly as an example.

For 1 and 2 above, the student is performing correctly; this is what the student is supposed to do. For 3 and 4, the student is making errors. Errors like 3 above are called "false negatives"--the student is mistaking examples for nonexamples. Errors like 4 are called "false positives"--the student is mistaking nonexamples for examples.

To construct a test that is diagnostic, it is necessary to construct items so that it is possible to tell whether the student is performing 1 and 2 correctly, and to determine why the student might be doing 3 or 4. To do this, the test is constructed systematically from the critical and variable characteristics.

For 1 and 3 above, the test must include at least one item that has all critical characteristics. This item will be a "true example."

For 2 and 4 above, the test must include items with critical characteristics systematically deleted. Suppose we have four critical characteristics. Then, we will need four items, each with a different critical characteristic deleted.

In general, if we have "n" critical characteristics, we will need at least "n + 1" items; one with all criticals present and the rest with different criticals deleted. This is the minimum number of items needed in order to be diagnostic. An excellent method for specifying the types of items required for a category is to construct a chart that lists the number of items required (n + 1) and each characteristic. A chart for a category with 3 characteristics (A, B, and C) would look like the following:

SAMPLE CHART FOR A CATEGORY WITH THREE CHARACTERISTICS

	Characteristic A		Characteristic B		Characteristic C	
Item #	yes	no	yes	no	yes	no
1	x		x		x	
2	x			x	x	
3	x		x			x
4		x	x		x	

After the minimum number of items have been determined, items for exceptions and variable characteristics that lead to common errors must be specified. Variable characteristics lead to errors because students treat them as if they were critical. For example, a student may think that a variable characteristic MUST be present when it doesn't have to be. This will lead to errors like 3 or 4 above. A "true example" that differs on a variable characteristic might be called a nonexample (3), or a nonexample that has a particular variable characteristic might be incorrectly called an example (4). Therefore, it is important to identify the variable characteristics that students might think are critical.

To construct items that include variable characteristics, we will use the "n+1" items that test the criticals and vary their variable characteristics to construct other new items. That is, from each of the "n+1" items, we will build other items that differ on variable characteristics. Exactly how many items are needed depends on what variable characteristics lead to common errors.

Let's determine what types of items are needed for our "table" example. Our "table" category has three critical characteristics: (1) it has a flat surface, (2) the surface is horizontal, and (3) the surface is supported from the floor. Therefore, we will need at least four test items. First, we will have one that has all critical characteristics--that is, a real table. Second, we need one that does not have a flat surface but that is horizontal and is supported from the floor--for example, a kitchen sink. Third, we need one that has a flat surface that is supported from the floor, but that is not horizontal--for example, an easel. Finally, we need one which has a flat horizontal surface but that is not supported from the floor--for example, a shelf. A chart that specifies these four items for this category would look like this:

CHART FOR TABLE CATEGORY

Item #	Flat Surface		Horizontal Surface		Support from Floor	
	yes	no	yes	no	yes	no
1	x		x		x	
2	x			x	x	
3	x		x			x
4		x	x		x	

Next, we need items corresponding to our exceptions. In this case, we need a "drafting table" exception. Finally, we need to consider which of the variable characteristics might be confused for critical ones. In the "table" example, for a young child just learning about tables, any of the variable characteristics we identified might be confusing. For example, a child might think, "if I can sit on it, it's not a table," or that all tables must be rectangular or that they must be made of formica. Therefore, we would need to vary our four sample items, to make other example and nonexample items that vary in height, shape, and material. Exactly how many items are needed depends on a common error analysis that would specify which of the variable characteristics lead to confusion.

At this point, we have identified the types of items that are required for the test to be diagnostic. It should be emphasized that the assistance of a subject matter expert will be necessary in analyzing critical and variable attributes for categories, particularly if the categories are not well defined in training objectives.

Step 2. Determine Item Formats

Step 2A. Determine "Response" Requirements. The training objective for a category task should specify the "response" the student is expected to make, and this exp·cted responsє will determine the allowable item formats. As stated above, USE-CATEGORY objectives typically require the student to classify, categorize, sort, or identify given instances according to type. These objectives usually require the student either to (1) sort or categorize or classify given instances into given categories, or (2) to give the category name for given instances. If the categories or category names are given, this is essentially a multiple-choice or matching situation. On the other hand, if the category names must be produced, constructed-response items are most appropriate.

Unfortunately, many training objectives fail to specify which type of response is required. For example, an objective might just say "the student will identify jet pumps." In these cases, the key to determining item formats is to REMEMBER THE JOB; that is, to determine whether later job performance will require the student to use given categories or to produce category names.

Another "response" decision to be made concerns what the student will do physically. For example, the objective, "the student will sort given metal fasteners according to type (bolts, screws, studs, nuts, rivets)," seems to require that the student actually sort given fasteners into bins. In contrast, the objective, "the student will identify the type of jamming appearing on an electronic warfare scope display," requires the student to state or write the type of jamming.

Step 2B. Determine Testing Conditions. The critical and variable characteristics determined above are also useful in identifying how examples and nonexamples must be presented to the student for classification. Since the critical characteristics are used by the student to make classification decisions, the conditions must be designed so that the student can perceive the critical characteristics. Confusable variable characteristics must also be perceptible.

For example, suppose an objective requires a student to classify something in which motion is critical. In this case, a written printed test with photographs cannot be used, because the time-varying or moving aspects of the displays cannot be presented.

Step 2C. Select Item Format. After determining the response required by the objective and the testing conditions needed to present the critical and variable characteristics, choosing the item format should be easy. If you have difficulty review what you did in steps 2A and 2B. Sometimes, constraints on testing may require that selected-response formats be used instead of appropriate constructed-response items. In this case, it is important to document the fact that an inconsistent item format was used.

Step 3. Construct Actual Test Items

Step 3A. Determine Actual Number of Items Required. The types of items specified in step 1 cover all the critical ard variable characteristics that are likely to lead to errors. However, we have not decided how many items of each type will be needed on the tests. When the test is actually given, will we need more than one? The answer to this question is usually "yes" for several reasons:

1. Often parallel forms of tests are needed for test security.

2. During formative evaluation, one wants to identify items that are unusually hard or easy or are strange in some other way, compared to other items. The more items you have, the easier it is to detect strange items and eliminate them.

3. More items increase the "reliability" of measurement. These evaluation questions are addressed in more detail in Chapter 6.

Step 3B. Build Items. Put the item types determined in step 1 into the formats determined in step 2. Develop additional items for each of the item types, but which vary on variable characteristics.

Step 4. Develop Standards and Instructions for Scoring and Diagnosis

Step 4A. Setting Standards. As stated earlier, standards for USE-CATEGORY tasks are determined by the requirements of the objective. For most categorization tasks, the objective will require complete accuracy. Some categorization tasks, however, cannot be performed perfectly, even by job experts, because performance is limited by some aspects of the task. For example, some radar or sonar detection and classification tasks are "noisy," so that it is difficult to perceive critical characteristics. Another example of this type of task is determining whether or not an electro-cardiogram indicates the presence or absence of heart disease. Here, actual EKGs are not "clean," and it can be difficult to detect critical characteristics. In these cases, standards should be set as high as possible but not so high that even experienced job performers fail. During training, a good strategy is to simplify early training and tests by removing as much "noise" as possible so that critical characteristics are clear. Final tests, of course, must be job-like.

Step 4B. Developing Instructions for Scoring and Diagnosis. As indicated in earlier chapters, it is important to develop clear instructions to test administrators about how to give and score the tests. These should include directions concerning what equipment must be available, time limits, personnel requirements, etc.

Because test items are constructed systematically, diagnosis should be straightforward. However, the diagnostic plan must be explained clearly so people who will actually use the test can prescribe remediation. This means that explanations and examples of how to interpret particular patterns of errors must be given.

Step 5. Develop REMEMBER-level Test Items if Necessary. If REMEMBER-level objectives will be tested during the USE-CATEGORY test, develop the necessary test items using the procedures described in Chapter 3.

REMEMBER-level test items may be given orally or on a written test and may be administered during or prior to the USE-CATEGORY test. REMEMBER-level test items should not be given during the USE-CATEGORY test if they interfere with the performance of the classification task. This could occur if a classification had to be made in a set amount of time or if several characteristics had to be evaluated at the same time.

REMEMBER-level items given during a USE-CATEGORY test should test memory for category characteristics, symbol names, and definitions of unfamiliar words and technical terms.

After items are developed, it is necessary to conduct tryouts, and revise the items if necessary. This process is described in Chapter 6.

EXAMPLE: Designing Test Items for a USE-CATEGORY Objective

Here is a USE-CATEGORY objective:

"Given any U. S. Navy CALL SIGN, the student will classify it according to one of the following types: International U.S. Navy Ship, International U.S. Navy Shore, Indefinite, Task Organization, or 'not a valid Navy call sign'. Type names will not be provided during classification."

Step 1. Analyze the Objective. This objective includes four categories: ship, shore, indefinite, and task organization. The definitions of these types are given in Navy publications:

1. International Ship: Starts with "N." Has three more letters (not numbers) after the "N."

2. International Shore: Starts with "N." Has two more letters after the "N."

3. Indefinite: Starts with "N." Has one more letter after the "N," except that "R" cannot be used.

4. Task Organization: Must consist of "Letter--Digit--Letter--Letter." (Any letters can be used.)

These definitions include the critical characteristics for each type of call sign, but they need a little work so that the critical characteristics are clear. (See Step 1A.)

1. International Ship:	(1)	Starts with "N."
	(2)	Three more characters after "N."
	(3)	All characters must be letters.
2. International Shore:	(1)	Starts with "N."
	(2)	Two more characters.
	(3)	All characters must be letters.
3. Indefinite:	(1)	Starts with "N."
	(2)	One more character.
	(3)	Character must be a letter.
		(Exception--"NR" cannot be used.)
4. Task Organization:	(1)	Four characters total.
	(2)	Second character must be digit
	(3)	Other characters must be letters.

The definitions above also include an exception: "NR" cannot be used as an indefinite call sign. (See Step 1B)

The next step (see Step 1C) is to list the variable characteristics. In this case, it does not matter which letters or digits are used so these are variable. Also, it does not matter whether pronounceable words result, so this is also variable (there is another type of call sign which uses words--it is important not to confuse these). Therefore, our variable characteristics and their values are:

1.	Letters:	A - Z
2.	Digits:	0 - 9
3.	Pronounceable?	YES or NO

The next step (see Step 1D) is to use the critical and variable characteristics to specify types of items necessary to test the objective. In this case, the situation is slightly complicated because we are dealing with four call signs instead of just one. Further, some of the call signs have the same critical characteristic. Therefore, let's specify item types one category at a time. First, we will build sample items from the item types determined from the critical characteristics for each category. Then we will examine the variable characteristics to determine what additional items are required.

1. International Ship. This call sign has three critical characteristics, so we will need four sample (n=3 + 1) items. We will need one item that starts with "N" and has three more letters--this is our true example. Next, we need an item that starts with "N," but that has the wrong number of letters following the "N." Then, we need an item that starts with "N," but has three characters some of which are numbers instead of letters. Finally, we need an item that has the right number of letters, but doesn't start with "N." These four items are the samples for International Ship call signs. A chart for the International Ship category would look like this:

CHART FOR INTERNATIONAL SHIP CATEGORY

Item	Starts with N		3 Characters after N		All Letters	
#	yes	no	yes	no	yes	no
1	x		x		x	
2	x			x	x	
3	x		x			x
4		x	x		x	

2. International Shore. This call sign also has three critical characteristics so four sample items are required. We need one item that starts with "N" and has two more letters (true example), one item that starts with "N" but with the wrong number of additional letters, one item that starts with "N" with two characters some of which are numbers, and one item that has the right number of letters but doesn't start with "N." A chart for International Shore would be similar to the International Ship chart.

3. Indefinite. There are again three characteristics, and four sample items are required. We need one item that starts with "N" and has one more letter (not "R") (true example), one item that starts with "N" and has one more character that is not a letter, one item that has two letters but does not start with "N," and one item that starts with "N" and is followed by the wrong number of letters. A chart for Indefinite would be similar to the International Ship chart. Finally, because "NR" is an exception we also need an "NR" item.

4. Task Organization. Again, this call sign has three characteristics and requires four sample items. We need one item that has four characters with the second character a number and the other characters letters (true example). The remaining items should include one item that has the wrong number of characters with the second character a number, one item that has four characters with the second character a letter, and one item that has four characters with the second character a number and some of the other characters numbers. A chart of Task Organization would be similar to the International Ship chart.

> Note. In this example, we have 17 sample items. However, we don't really need that many in this case. The reason is that we are dealing here with more than one category, and examples for one category can serve as nonexamples for another. For instance, the "true example" for International Ship call signs--"N" + 3 letters--is also a nonexample for all the other categories. Situations like this with several related categories are called "coordinate categories." In these situations, the sample items should be checked, and duplicates should be deleted.

The next part of Step 1D is to consider the variable characteristics. In this case, the letters and digits can and should vary, and we should include both pronounceable call signs and call signs that are not pronounceable. Our complete list of sample items is as follows:

	ITEM	WHAT CHARACTERISTICS?
1.	NPKC	Example of International Ship, not pronounceable. Nonexample of Int. Shore, Indefinite, Task Organization.
2.	NEWS	Same as above, except pronounceable.
3.	NIT	Example of International Shore. Nonexample of Int. Ship, Indefinite, Task Organization. Pronounceable.
4.	NGH	Same, except not pronounceable.
5.	NT	Example of Indefinite. Nonexample of Int. Ship, Int. Shore, Task Organization.
6.	NR	This is our exception.
7.	NA55	Nonexamples with digits.
8.	NY8	
9.	N7	
10.	AGKL	Nonexamples--don't start with "N."
11.	ROW	
12.	TB	
13.	NFHDW	Nonexample--too many letters.
14.	K3XC	Example of Task Organization. Nonexample of the others.
15.	N4MW	Same as above. (Note possible confusion with Int. Ship.)
16.	P3BGR	Nonexample--too many letters.
17.	WH9S	Nonexample--digit in wrong position.

Notice that, in these examples and nonexamples, we've been careful to vary the letters and digits.

At this point, we have sample examples and nonexamples that could be used in actual test items. We still have to determine item formats and the number of items needed.

Step 2. Determine Item Formats. This objective requires the student to classify given call signs. The objective further states that type names will not be given during classification, so this means that multiple-choice or matching formats cannot be used. Therefore, we should use a "fill-in" format, which requires the student to write or state orally the category name for each given call sign on the test.

The test items will have to be presented so that the student will be able to identify the critical and variable characteristics. This means that we probably ought to present the call signs in print.

Step 3. Construct Actual Test Items. First (see Step 3A), it is necessary to determine how many test items will be necessary. In this case, let's assume that we will need at least two "parallel" forms of the test for test security. Also, it would be better to have additional items in case some are found to be strange during tryout. So, four or five additional sets of items should be built. In this example, it is simple to build more items: We simply take each of our 17 sample items, and write others that look the same but with the variable letters or digits charged.

A more difficult question involves determining how many items must be included on the final version of the tests so that accurate information about student competence can be obtained. Do we need just 17, 34, or more? A final answer to this question must wait until tryouts of the test have been conducted. However, because the sample items have been constructed systematically, 17 items should be enough to give good diagnostic information. Also, a subject matter expert could be consulted to determine how many items are needed to give accurate information about student competence.

Second (see Step 3B), we put the sample items into the format determined in Step 2. In this case, this just means that we list the items, with instructions to the student to "write the type of each call sign given below in the space provided...." The instructions should also tell the student what response to make to nonexamples; that is, invalid call signs: "for call signs which are NOT valid Navy call signs, write 'invalid.'" Naturally we would scramble the order of the items.

Step 4. Develop Standards and Instructions for Scoring and Diagnosis. What standard should be set for the call signs test (see Step 4A)? The objective implicitly specifies that the task be performed with complete accuracy. There seems to be no reason to change this standard. Again, the goal of tests like this is diagnosis of the reasons for inadequate performance, and errors mean that the student needs remediation.

Instructions about how to administer the test are straightforward, since no special equipment, performance monitoring, etc. are required.

Instructions for scoring and diagnosis (see Step 4B) are needed so that test administrators can accurately interpret test results and plan remediation. To build these instructions, we need to list each of the items, possible responses the student could give, and possible diagnoses for each of the responses. This is done below for some of our sample items: (It is not necessary to do this in quite as much detail as we have done here, but enough detail is necessary to that different patterns of student responses can be traced to particular misconceptions.)

ITEM	POSSIBLE RESPONSE	POSSIBLE DIAGNOSIS
1. NPKC	Int. Ship	Correct
	Int. Shore	Student has Ship and Shore confused (see Int. Shore below to determine what the confusion is), OR the student knows Shore must start with "N" but doesn't know how many letters and doesn't know Ship. (See Int. Shore below to check for number of letters and see other Ship items to check knowledge of ship call signs.)
	Indefinite	Has Ship and Indefinite confused, OR knows Indef. must start with N but doesn't know how many letters,and doesn't know Ship. See Indef. below.

	Task Org.	Has Ship and Task Org. confused, OR doesn't know either Ship or Task Org.
	Invalid	Does not know at least one of the criticals for Int. Ship OR thinks that some variable is a critical. (See responses to other items below.)
	Other	Student misread instructions. Student may also not know Int. Ship. (See other responses below.)
2. NEWS	(Same as above)	(Same as above) If student was correct on this but put "invalid" on 1, then student thinks pronounceability is critical. Vice versa for nonpronounceable.
3. NIT	Int. Ship	Has Ship and Shore confused, OR knows Ship must start with N but doesn't know how many letters and doesn't know Shore. (See Ship above.)
	Int. Shore	Correct.
. (etc.)		
17. WH9S	Int. Ship	Does not know Ship must start with N and have no digits.
	Int. Shore	Does not know Shore must start with N and have no digits.
	Indefinite	Does not know Indef.
	Task Org.	Does not know that digit must be in second position.

Invalid	Correct
Other	Student misread instructions. (See responses to other items.)

It is important to notice several things about this list of diagnoses. First, you can't tell from any single item why a student may have made an error. Instead, a pattern of responses must be considered. For example, if a student makes a certain error on items 1 and 2 and makes other errors on later items, this tells us that the student has two of the types of call signs confused. Other patterns of errors may tell us that the student does not know the critical characteristics of some call sign type, etc. Sometimes, you will get an "inconsistent" pattern; that is, sometimes the student will be correct and other times incorrect, or the student will be incorrect in a variety of ways. This means either (1) you haven't identified all the variable characteristics that may lead to errors, or (2) the student doesn't know anything and is just guessing. If other students are doing all right, guessing is a more likely explanation.

Second, it is not enough just to score the student "correct" or "incorrect" on particular items. It is necessary to look at actual student answers, because wrong answers give information about what the student is thinking. Notice also that wrong answers often give information about some category other than the one in the item. It should be clear, of course, that it is meaningless just to say that a student got 80% (or whatever) on the test. This gives no information about what the student really knows (unless the student got 100%).

Step 5. Develop REMEMBER-level Test Items if Necessary. No REMEMBER-level objectives will be tested in this example.

Testing USE-RULE Objectives

As stated earlier, A USE-RULE objective requires the student to solve a large number of problems or perform a complicated series of steps on a large number of different objects, events, etc. Instead of having to remember each problem or the steps for each different object, the student is taught a rule that can be used to deal with problems not seen before.

As in other USE-level tasks, the main problem in constructing test items for USE-RULE tasks is analyzing the objective or task to determine what should be tested. RULE objectives should include a complete statement of the rule or reference to where a complete statement can be found. Sometimes there will be a good statement of a rule available in a technical manual, and occasionally there will be a complete flow chart for the rule process. In some cases, a subject matter expert will have to be consulted to obtain a complete statement of the rule.

Rules consist of sequences of actions or STEPS to be performed, and questions or DECISIONS to be answered. (In general, procedures just have steps. Rules, because they are meant to deal with a variety of situations, have decision points in them.) Often rule statements are written out "verbally," but for our purposes the best way to represent a sequence of steps and decisions is to draw a FLOW-CHART of them. In a flow chart, steps are usually put in rectangular boxes:

```
 ___________________
|                   |
| Step 2. ...       |
|                   |
|___________________|
```

Decisions are usually put in diamond-shaped boxes, with a question inside and possible answers to the question outside:

```
          |
         / \
        /   \
       /     \
      /  Is   \
     / this a  \____
     \ DIAMOND /  no
      \   ?   /
       \     /
        \   /
         \ /
          |yes
          |
```

Usually, decision questions are "yes-no" type questions, but they don't have to be; there can be three or four or more answers out of a decision diamond.

These flow charts are the same as those used in work with computers, as fault-logic diagrams for electronic troubleshooting, and in some technical manuals for rule processes. Two books that discuss how to do flow charting are The Instructional Design Library Vol.II: ALGORITHMS by Horabin and Lewis (1978) and Case Studies in the Use of Algorithms by Lewis (1978).

It is important to note that there is usually more than one correct way to flow chart a rule. What is important is to identify all the decisions and steps and arrange them correctly. The steps needed to analyze rules, and develop test items and diagnostic test are described below.

Step 1. Analyze the Objective to Determine the Test Item Domain.

The first step in constructing tests for RULE objectives involves developing a flow chart of the rule. This means that the verbal statements must be translated into steps and decisions. Steps are actions that must be performed, while decisions require that some question be answered. The answer to the question then leads to other steps or decisions. For example, suppose we had the following objective and rule statement:

Objective: "Given the weight of a thawed raw fowl, and depending on whether or not the fowl is to be stuffed, the student will determine the optimum cooking time at 325 degrees fahrenheit."

Rule statement:

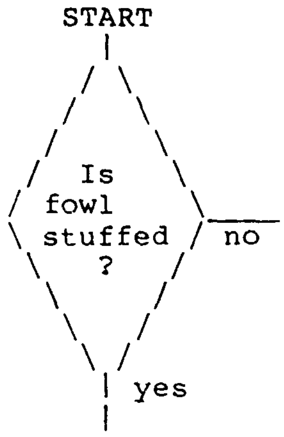
Unstuffed, weighing 6 lbs or under--Multiply the weight in pounds by 20 to find cooking time in minutes.

Unstuffed, weighing more than 6 lbs.--Multiply the weight by 15.

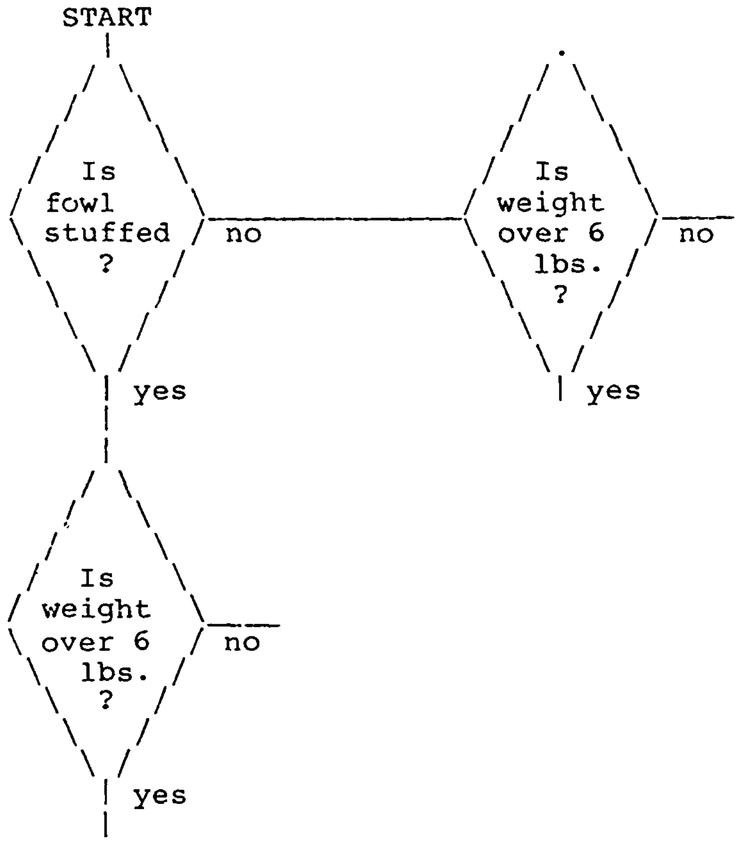
Stuffed, weighing 6 lbs. or under--Multiply weight by 25.

Stuffed, weighing more than 6 lbs.--Multiply weight by 20.

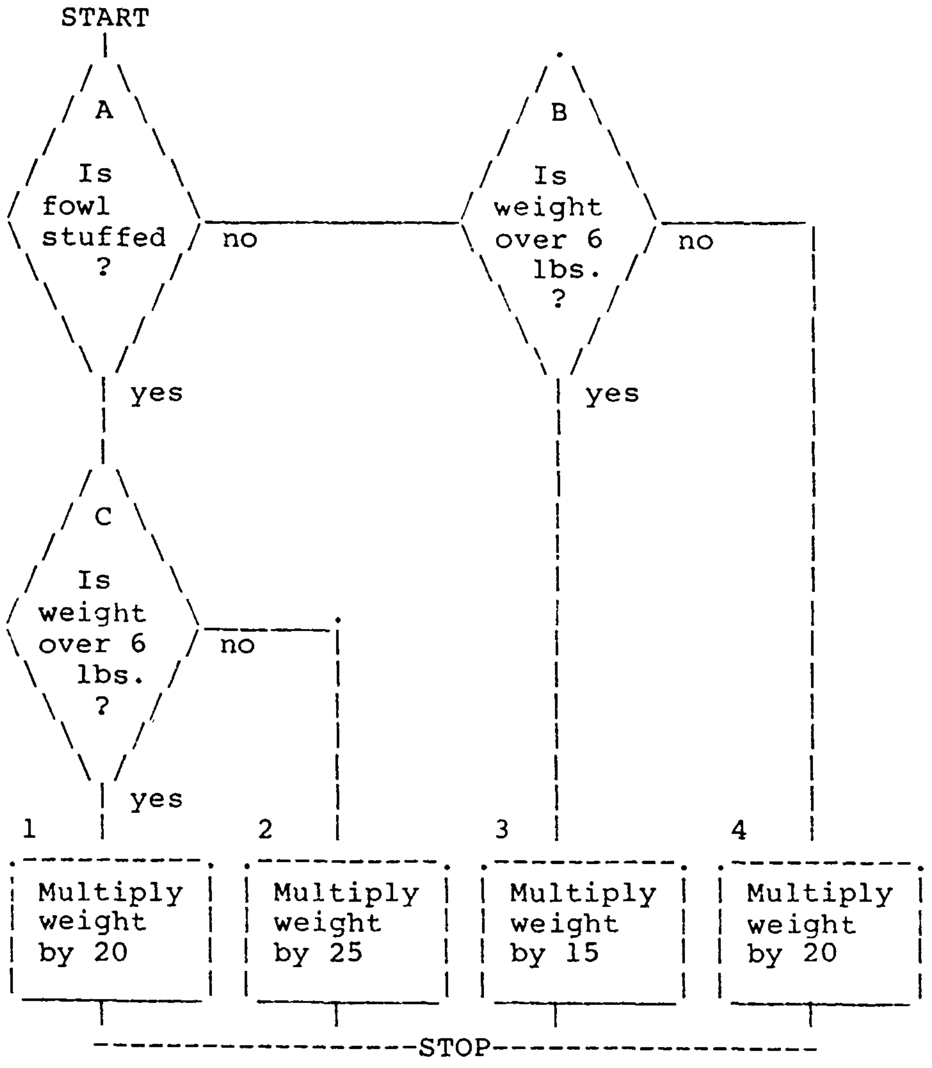
First, it is necessary to identify the decisions to be made in this rule. In this case, the student must first decide if the fowl is stuffed or not. Therefore, our flow chart will begin with:



Next, whether the fowl is stuffed or not, the student must determine whether the weight is 6 lbs or less, or more than 6 lbs. So, our flow chart becomes:



Once the student has determined whether or not the fowl is to be stuffed and the weight, the steps necessary to determine cooking time can be performed. These are put in rectangles in the flow chart:



(Note: In this chart, we've labeled diamonds with letters and steps with numbers for future discussion.)

Once the rule has been flow charted, we can specify how many types of items will be necessary to test a student's application of the rule.

To test a student's performance of the rule, we need to test all of the rule; that is, each step and decision. This also allows diagnosis. If each step and decision is tested, we can tell from a pattern of responses which decisions the student cannot make or which steps the student cannot perform. In general, we need enough items so that each different step or action in the flow chart is tested. (Sometimes we also need items to test decisions which may have been skipped).

For example, in the flow chart above, one problem might result in a "yes" for decision "A," and a "yes" for decision "C," leading to step 1. Another path might be a "no" on decision "A," and a "yes" on decision "B," leading to step 3. In this flow chart, there are four of these paths; that is, there are four different actions that should be tested.

More complicated flow charts may have many different patterns of actions and decisions. Some flow charts have "loops" or "branches" that split off and later rejoin some sequence of steps. In these cases, it is possible to build items that go through several of the loops or branches at once, so that many steps and decisions are tested with just a few items. Selecting which loops or branches should be tested in a single item depends on what is typically done when the task is performed on the job. An example of this is given later in this section.

There is one other important issue that must be considered when developing test items for rules. It concerns the difference between decisions and steps. Decisions are made cognitively and therefore can usually be tested with a paper and pencil or oral test. Steps, on the other hand, are actions or sequences of actions and must usually be tested with a performance test. Remember that, since some action sequences are already paper and pencil (e. g., doing a calculation or filling out a form), the performance test is a paper and pencil test. The only time steps do not need to be tested is when the actions are so simple or trivial that all students could be expected to do them without any training. The reason this is important for test development is that usually steps only need to be tested once. The problem is that, depending on the flow chart, one step may have to be tested several times to test all the decisions. This is often a waste of valuable test time. The solution is to test the performance of each step once and test the decisions using a paper and pencil test. A book that describes procedures for designing these types of test is Construction and Use of Written Simulations by McGuire, Solomon, and Bashook (1976).

The test items built from a flow chart will give information about whether or not a student can make the decisions or perform the steps of a rule. If it is necessary to get additional, more precise, diagnostic information about a particular decision or step, then more analysis and testing are necessary. To do this, a decision should be analyzed as if it were a USE-CATEGORY task and a step should be analyzed as if it were a USE-PROCEDURE task.

Step 2. Determine Item Formats

Determining response requirements and testing conditions is the same as for PROCEDURES and CATEGORIES. The student must be asked to perform whatever is required in the objective. If the decisions in the rule are easy to make and most of the difficulty is in performance of steps, the test ought to use the real task and equipment. If, on the other hand, the performance steps are easy or have been taught for previous objectives but the decisions are difficult to make in the right order, simulations should be used.

In cases where the rule results in something that can be written, like calculating a value (as in the fowl example), fill-in tests are recommended. Multiple-choice tests can also be used, but these are not the best choice since students can use the alternatives to help guess the correct answer. If multiple-choice tests must be used, the best way to construct items is to use the answers from paths other than the correct one as the alternatives. If there are a large number of paths or if the rule is complicated, alternatives for multiple-choice items can be generated by giving students who have been through the training problems to solve and using their incorrect answers for alternatives. In the fowl example, we traced through incorrect paths to develop items with alternatives like the following:

"If you have a 12-pound stuffed fowl, it should be cooked at 325 degrees for how long?

a. 240 minutes

b. 180 minutes

c. 300 minutes"

Notice here that each of the wrong answers (b and c) comes from using a wrong path through the flow chart.

Of course multiple-choice tests don't work for rules that require performing some physical action. For these tasks, it is necessary to determine whether product or process measurement is to be used and to construct checklists and/or rating scales. The same process as in Chapter 4 for USE-PROCEDURE tasks should be used. Product measurement should be used only rarely for USE-RULE tasks since, in order to be diagnostic, the process must usually be observed.

Step 3. Construct Actual Test Items

Obviously, once the flow chart paths are identified, it is easy to tell what the content of the different types of items should be. Just go through the flow chart, building items that trace out as many actions as possible. Each action must be tested at least once. In our fowl example, we need at least one item for each path, so we need one item that asks the student to calculate cooking time fcr a stuffed fowl over 6 lbs., ore for a stuffed fowl under 6 lbs., one for an unstuffed fowl over 6 lbs., and one for an unstuffed fowl under 6 lbs. In the actual items, we should pick fowl weights that are typical of the values actually encountered on the job.

We again have the problem of determining how many items of each type must actually be developed. Just as in CATEGORIES, we will probably need additional "parallel" forms of tests and additional items in case some are "strange" during tryout. Again, the more difficult question is how many items of each type a student must do so that performance can be assessed accurately. The answer to this question depends partly on the results of tryouts (discussed in chapter 6), but a subject matter expert can also be consulted. Additional items for each type can be constructed simply by varying the given information (e.g., the weight of the fowl) in each type. For some rules, the number of items can be reduced by designing comprehensive items that test several decisions and steps at the same time. The book by McGuire, Solomon, and Bashook (1976) that was mentioned earlier in this chapter gives guidelines for developing these types of items.

Example items for oui fowl rule are:

"Calculate the cooking time in minutes at 325 degrees fahrenheit for each of the following fowl:

A. stuffed, weight is 3.2 lbs.

B. stuffed, weight is 14 lbs.

C. unstuffed, weight is 5.3 lbs.

D. unstuffed, weight is 18.7 lbs.

More items can be constructed just by varying the weights. Since this example is very simple, four items--one for each type--are probably enough.

Step 4. Develop Standards and Instructions for Scoring and Diagnosis.

Step 4A. Setting Standards. Accuracy standards for the decision making part of USE-RULE tasks should always be complete accuracy because of the way in which a rule is defined. In order for a task to be a USE-RULE, all the decisions must be able to be made with complete accuracy. There is no room for "noise" in a rule. If a decision can be made accurately only part of the time, then the task is USE-PRINCIPLE or USE-CATEGORY. The standards for performing the steps of a rule should also be what is specified in the objective. There may be time or tolerance standards that are applicable to the rule steps. These are determined by the requirements of the job being taught and should be specified in the objective. These standards should be incorporated in the rating scales or checklists used for individual steps (See Chapter 4).

Step 4B. Developing Instructions for Scoring and Diagnosis. See instructions for Step 4B, for USE-CATEGORY tasks. In the next section, we will work through an example that illustrates the process of developing tests for USE-RULE tasks.

Step 5. Develop REMEMBER-level Test Items if Necessary. If REMEMBER-level objectives will be tested during the USE-RULE test, develop the necessary test items using the procedures described in Chapter 3.

REMEMBER-level test items may be given orally or on a written test and may be administered during or prior to the USE-RULE test. REMEMBER-level test items should not be given during the USE-RULE test if they interfere with the performance of the USE-RULE task. This could occur if the rule had to be performed in a set amount of time.

REMEMBER-level items given during a USE-RULE test should test memory for rule steps and decisions, symbol names, and definitions of unfamiliar words and technical terms.

EXAMPLE: Designing Test Items for a USE-RULE Objective

Here is a USE-RULE objective:

"Given an automobile that is reported to have a dieseling problem (engine continues running after turn-off), the student will diagnose and repair the problem."

Rule Statement:

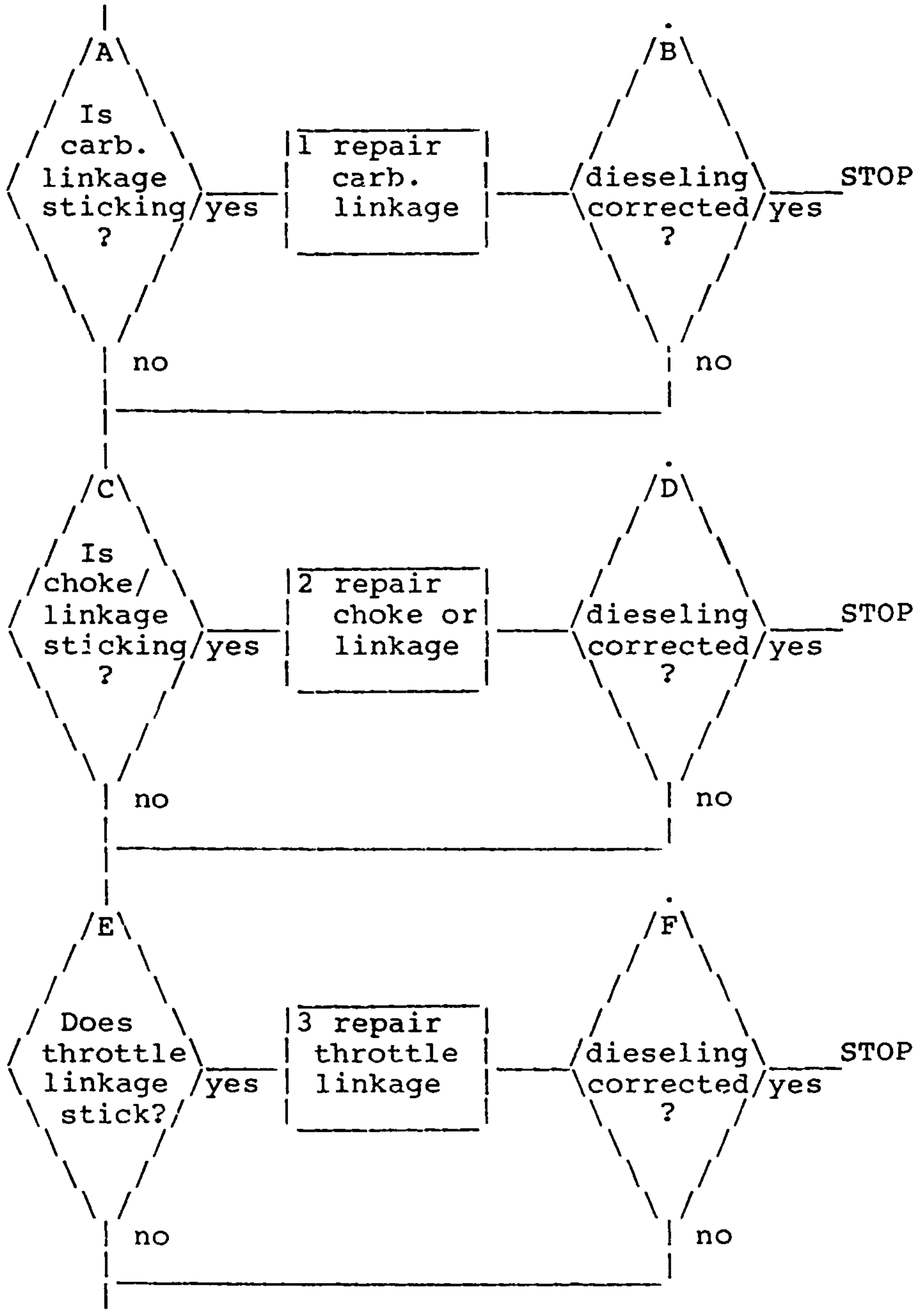
1. Check carburetor linkage, choke and linkage, throttle linkage, and fast-idle cam for sticking. Repair as necessary.

2. If dieseling persists, disconnect and plug vacuum advance hose, and reset dwell, timing, and idle RPM to tune-up specifications.

3. If dieseling persists after resetting dwell, timing, and RPM, disconnect idle solenoid wire to check for RPM drop. If there is no RPM drop, replace solenoid. If there is an RPM drop, go to step 4.

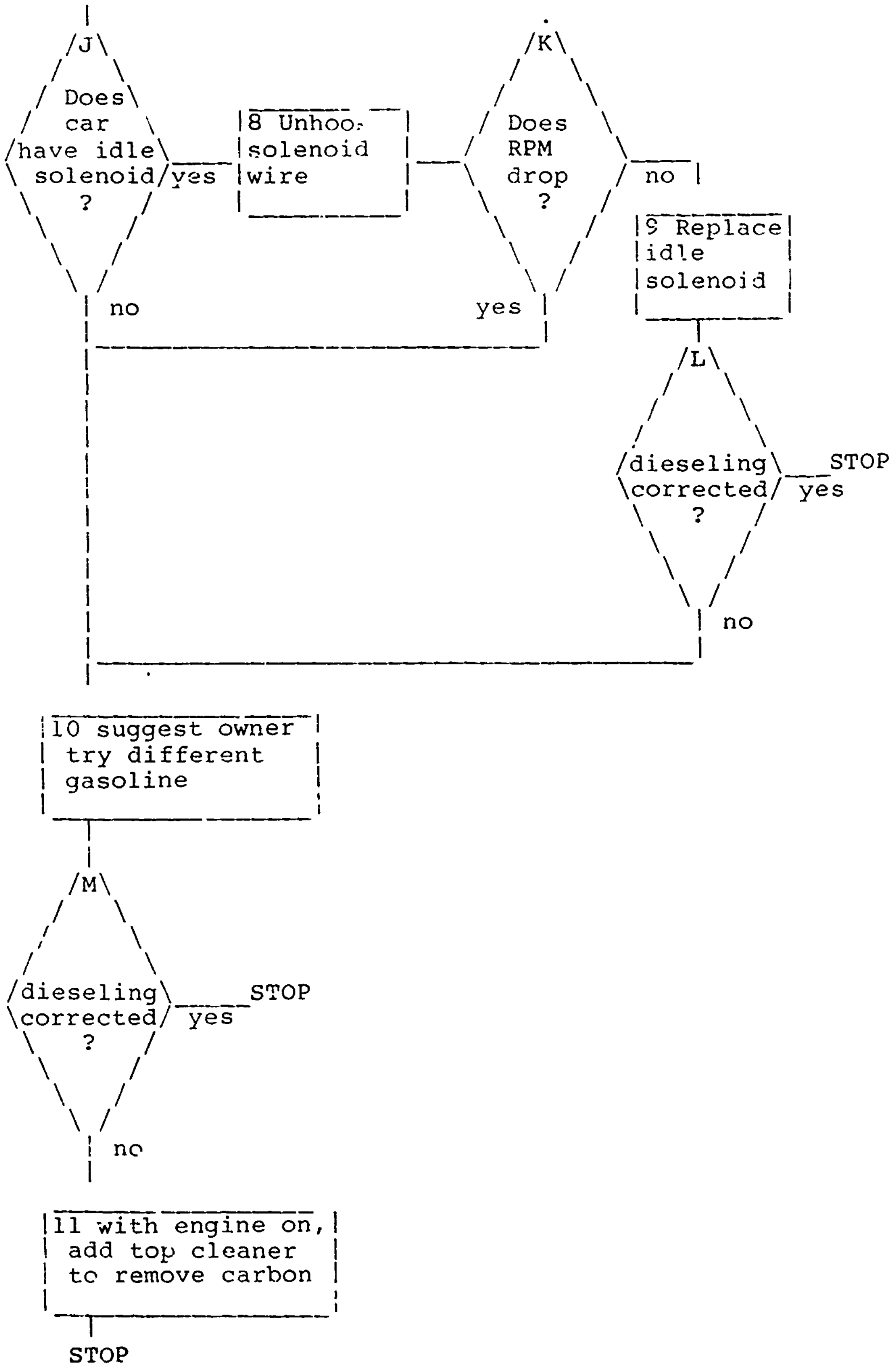
4. Use a higher octane gasoline. If condition still exists, add top engine cleaner through carburetor to remove carbon.

Step 1. Analyze objective to determine test item domain. The flow chart for this rule is shown below. We have put letters in each of the decisions and numbers for the steps so we can refer to them later.

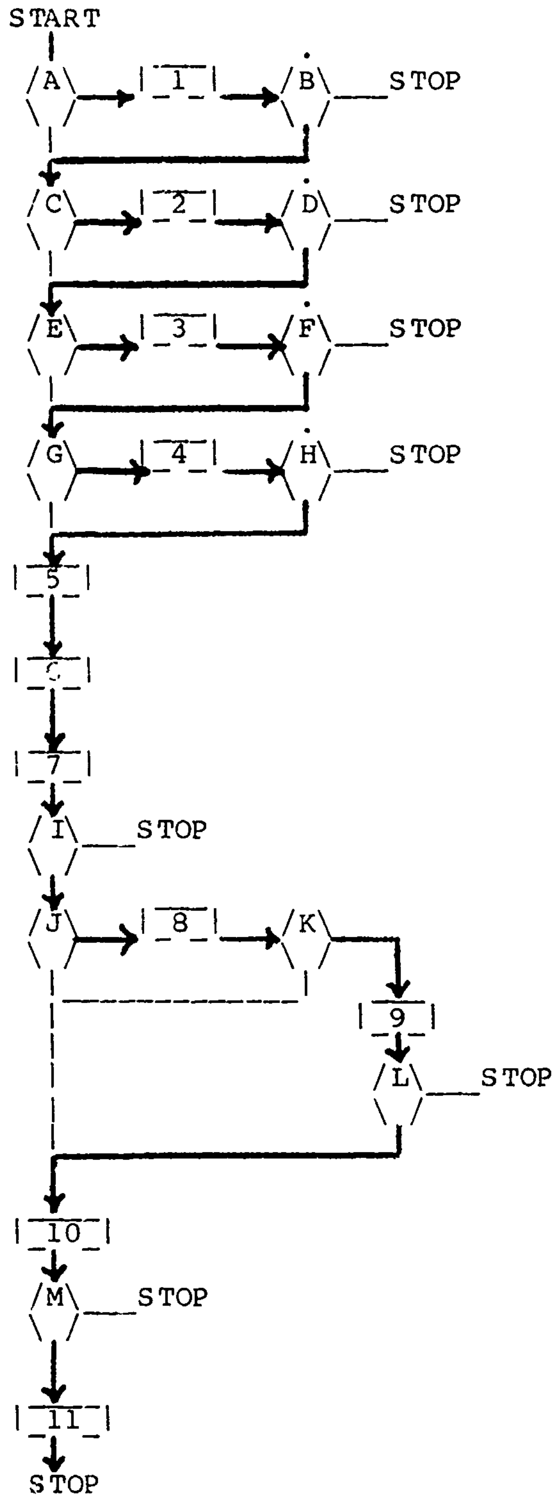

A
Is carb. linkage sticking ?
yes
1 repair carb. linkage
B
dieseling corrected ?
yes
STOP
no
no
C
Is choke/ linkage sticking ?
yes
2 repair choke or linkage
D
dieseling corrected ?
yes
STOP
no
no
E
Does throttle linkage stick?
yes
3 repair throttle linkage
F
dieseling corrected ?
yes
STOP
no
no

(next page)

```
          |                                     .
         /G\                                   /H\
        /   \                                 /   \
       / Is  \                               /     \
      / fast- \      |4 repair   |          /       \
     ( idle cam\_____| fast-idle |_____/dieseling\____STOP
     \ sticking/yes  | cam       |     \corrected/yes
      \   ?   /      |___________|      \   ?   /
       \     /                           \     /
        \   /                             \   /
         \ /                               \ /
          | no                              | no
          |                                 |
          |_________________________________|
          |
          |
   _________________
  |5 disconnect and |
  | plug vacuum     |
  | advance hose    |
  |_________________|
          |
   _________________
  |6 Connect tach-  |
  | dwell meter and |
  | timing light    |
  |_________________|
          |
   _________________
  |7 set dwell,     |
  | timing, and RPM |
  | to specs        |
  |_________________|
          |
          |
         /I\
        /   \
       /     \
      /       \
     /dieseling\_____STOP
     \corrected/ yes
      \   ?   /
       \     /
        \   /
         \ /
          | no
          |
     (next page)
```


J
Does car have idle solenoid ?
yes
8 Unhoo: solenoid wire
K
Does RPM drop ?
no
9 Replace idle solenoid
no
yes
L
dieseling corrected ?
STOP
yes
no
10 suggest owner try different gasoline
M
dieseling corrected ?
STOP
yes
no
11 with engine on, add top cleaner to remove carbon
STOP

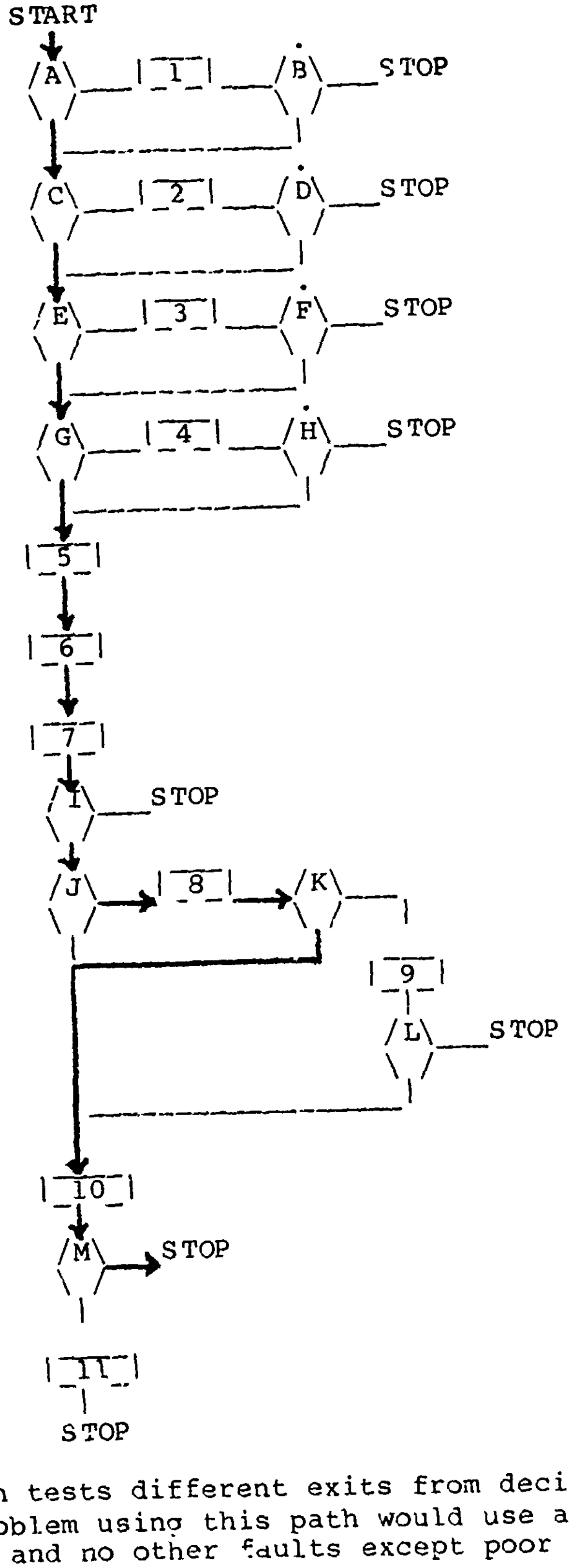
One path through the rule above will pick up many of the action steps. This path is shown below:



The flow chart path on the previous page translates into a problem requiring a car with all linkages sticking; RPM, dwell, and timing out of adjustment; a bad idle solenoid; cheap gasoline; and a dirty top engine. To troubleshoot the car, the student would have to perform all the steps in the flow chart.

At this point, one question to ask is whether the resulting type of item is job-like. Is it common on the job for a car to have all these problems at once? If not, this problem can be broken into two or more different problems that test different parts of the flow chart. For example, one problem might use a car with dirty linkages but no other problems. Another problem might have a bad idle solenoid only, etc. The resulting item types ought to be typical of the kinds of problems commonly encountered on the job.

We also need to test paths that use decision-answers not used in the first problem type. Another possible flow through the chart is shown on the following page:



This path tests different exits from decisions A, C, E, G, K, and M. A problem using this path would use a car with a good idle solenoid and no other faults except poor gasoline.

Step 2. Determine Item Format. This objective requires the student actually to perform the diagnosis and repair. This means that a performance test is necessary so that each of the steps can be tested at least once. Once each step had been tested, written simulations could be used to test the decisions. Finally, it appears that the steps of this task would be difficult to simulate. It is probably best to use real cars to test performance of the action steps.

Another decision to be made is whether to use rating scales or checklists for each of the performance steps. In this case, it appears that checklists are adequate for each of the actions. More complicated actions might require rating scales--see the valve example in Chapter 4.

A third decision is whether to use a written simulation to test the decisions. The book by McGuire, Solomon and Bashook (1976) that was mentioned earlier in this chapter gives guidelines for developing these types of items.

Step 3. Construct Actual Test Items. First, it is necessary to determine how many test items are required. In step 1, we chose different paths through the flow chart to serve as types of items to be tested. These paths specified the types of malfunctions to be used in test items. This task is a rule task in part because it applies to a variety of cars, not just a single model. Therefore, it is necessary to use problems on different makes (e.g. Fords, GM cars, Chrysler cars, etc.) and different engine models that have different linkages, are with and without idle solenoids, etc. Enough items of each type must be constructed so that we can tell if the student can perform each of the steps and decisions on a variety of makes of cars.

Step 4. Develop Standards and Instructions for Scoring and Diagnosis. The standard for this objective ought to be that the student performs all decisions and steps with complete accuracy. Because performance will actually be observed and checklists will be used, diagnosis by an instructor should be relatively easy. If a student fails to recognize a sticky linkage or performs a step incorrectly, the instructor can recognize this and give the student remediation.

Step 5. Develop REMEMBER-level Test Items if Necessary. No REMEMBER-level objectives will be tested in this example.

Testing USE-PRINCIPLE Objectives

As stated earlier, a USE-PRINCIPLE objective requires explanation, prediction, or diagnosis of a large number of possible situations, events, effects, problems, etc. Instead of having to remember each possible situation or event and its effects, the student is taught a "principle" that explains "how" or "why" situations or events occur. The student can use the principle to explain, or predict, or diagnose a variety of situations not seen before. Principles are somewhat like rules but are much more complicated and the analysis is more difficult and not as straightforward as the analysis for rules.

A variety of USE-PRINCIPLE tasks occur in Navy training. These tasks are usually of the following types:

1. Explanation. The student is asked to give an explanation of how a particular system operates. Note that this is a USE-level task only when the student has not previously been given the explanation. Real explanation tasks occur infrequently because a student is usually taught how a system operates at the REMEMBER-level. The student uses this explanation on the job to diagnose or predict. An example of a true explanation task would be if an electronics technician were given the schematic for an unfamiliar piece of equipment and asked to explain how it worked.

2. Prediction. The student is given some initial "boundary" conditions or initial assumptions and is asked to predict what is likely to occur as a result. These are often "what would happen if..." questions. Examples of prediction tasks are:

a. Given a particular pattern of weather conditions, predict the weather which will occur later.

b. Given an initial tactical situation, predict what is likely to occur.

Note that these are USE-level tasks only when the student has not previously been given the same problem either as practice or as an example.

3. Troubleshooting or Diagnosis. The student is given a particular pattern of symptoms and is asked to determine what is causing them. This diagnosis process is similar whether it involves a technician trying to locate a malfunction in an electro mechanical system, a doctor trying to determine the reason for an illness, or an instructor trying to determine why a student is missing certain test items. Again, these are USE-level tasks only when they are new problems the student has not seen before. Also, troubleshooting tasks are PRINCIPLES only when the fault-isolation process has not been flow-charted or "proceduralized." If the process has been flow-charted, it is a rule or procedure (see the example in the previous section). In

these cases, the student can do troubleshooting without really "understanding" the system's operat.ıon.

Although these PRINCIPLE tasks do not seem similar on the surface, the same process underlies all of them. When people think about complicated systems, they use "mental models." These models are more or less simplified descriptions of how the system works. These models are often "qualitative" rather than "quantitative"; that is, they describe the system in terms of the kinds of interactions that happen in the system, rather than calculating exact values. For example, people think about the braking system on their car in terms of force on the pedal being translated into force on the brake shoes--not in terms of exact measurements of pressure in the hydraulic fluid.

Mental models are used in all the PRINCIPLE tasks described above. In "explanation" tasks, the student has to build a model of the system and verbalize it. In "prediction" tasks, the student has to run initial conditions through the model to predict what happens. In "diagnosis" tasks, the student has to run the model "in reverse," from symptoms or effects backwards to possible causes.

All this means that USE-PRINCIPLE tasks must be analyzed in terms of the models people are likely to use in performing the tasks, so that good diagnostic tests can be constructed. The problem is to specify the models.

Step 1. Analyze the Objective to Determine the Test Item Domain.

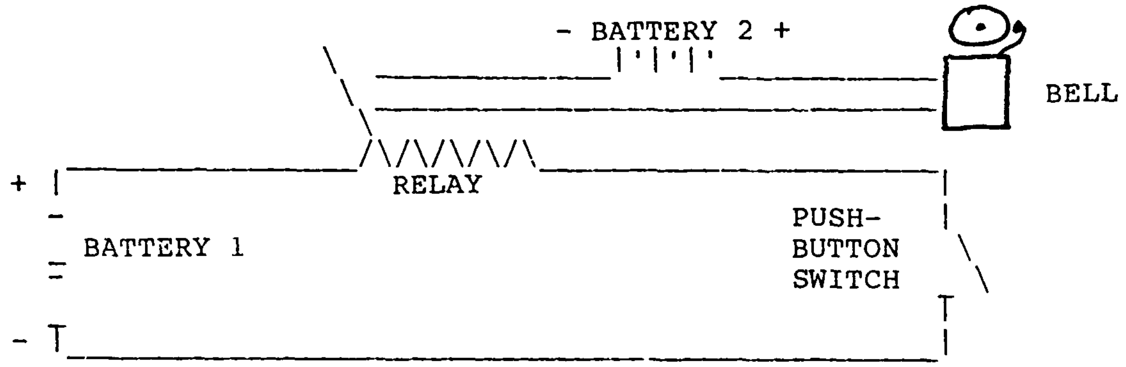
The first step involves obtaining a complete statement of the principle. Often the objective must be clarified by specifying exactly what situations, events, equipment, etc. it is meant to apply to. Then, we construct an explanation from the objective and the principle(s) to be used in satisfying the objective, either for each of the situations, events, or equipments specified or, if they are similar enough, for some "general" situation, event, or equipment. This explanation should include a description of the operation of the system and des criptions of the cause-effect relationships involved in ach. ving the objective.

For example, suppose we had the following objective:

"Given a schematic of an Alarm and Warning System D.C. circuit, the student will explain how the circuit operates, predict the effects of shorts or breaks in the circuit and diagnose where shorts or breaks have occurred to cause given symptoms."

This objective is only USE-PRINCIPLE if it means that students will be given new circuits not seen before during instruction. That is, the student is supposed to apply what he has been taught about how circuits work to explain what happens in new circuits. This knowledge is usually acquired during the course of instruction. On the test, the student must generalize this knowledge to new circuits. To clarify the objective, it is necessary to specify exactly what Alarm and Warning System D.C. circuits the student will work with. Let's make up one.

Here is a schematic, and an explanation of the operation of a circuit chosen for the student to analyze:



Explanation of Circuit Operation:

1. This circuit operates as follows: The push-button switch is normally open. In this condition, no current flows anywhere in the circuit. When the switch is closed, battery 1 supplies power to the relay, which operates and closes its contacts. These contacts complete another circuit so that battery 2 can power the bell.

2. Any short across the switch will cause the bell to ring; so will a short across the relay contacts. A short across either battery will cause that battery to discharge rapidly, and the circuit the battery powers will be inoperative.

3. A break in either circuit will prevent the bell from ringing when the switch is closed. If the break is in the bell circuit, the relay may still operate when the switch is closed. If the break is in the switch-relay circuit, nothing will happen.

The above explanation covers all the cause-effect relationships required by the objective. Notice that no mention is made of "quantitative" circuit values, or Ohm's Law, or the resistance of the relay coil, etc. The explanation is "qualitative" in that it describes what happens in the circuit in terms of what actions cause what effects on various components.

Often, explanations like this can be obtained from a technical manual or other publication. If an explanation must be constructed, one way is to start with whatever the "inputs" to the system are and then trace the effects through to the "outputs" of the system. Another way is to describe the flow of material, information, or actions through a system. A third way is to describe how some factors continuously interact or affect each other simultaneously.

One problem in constructing explanations like this is the level of detail required. The explanation above could have been simpler--"the button rings the bell"--or it could have been much more detailed. For example,

> "When the switch is closed, electrons flow through the relay coil creating a magnetic force which operates on the armature of the relay to move it so that contacts attached to it can close.....(etc.)"

The level of detail required depends on the specificity of the objective, the nature of the job the student is being trained for, and the student's state of knowledge about the system being described. Often, a subject matter expert will have to be consulted.

Once the explanation is constructed, the next step is to rewrite each of the cause-effect relationships as single statements. In our circuit example, we would have the following:

1. When the switch is pushed, the circuit is completed.

2. In a complete circuit, current flows.

3. Current flow causes the relay to operate.

4. When the relay operates, its contacts close.

5. (etc.)

These explanation statements will serve as item types for explanation test items.

To build prediction item types, we vary the situation by giving new boundary conditions or initial states, and then go through the explanation of how the system works again to identify what changes occur. The actual prediction items will give the student the initial conditions and ask what the effects will be. If the student must also explain the reasons for his prediction, this amounts to another explanation task, which is built as described above. In our circuit example, a short or break is a different initial state of the system. For example, a short across the switch contacts also completes the circuit, current flows, the relay operates, etc.

To build diagnosis item types, we start from symptoms and work backwards through our explanation to identify possible causes. The first problem in building diagnosis or troubleshooting items is where the "symptoms" come from. In our simple circuit example, it is easy to identify them but, in a complicated system almost anything can go wrong. The best strategy is to consult a subject matter expert to determine what symptoms and faults are most often observed on the job or are most critical, and use these.

Troubleshooting items are more complicated to develop than prediction items, because there may be several possible causes for an observed symptom. In our circuit example, if the bell rings even when the button is not pushed, this works back to a short somewhere in the system, but we will need additional items to get the student to pinpoint the fault.

Step 2. Determine Item Formats

Three item formats should be considered for explanation, prediction, and diagnosis types of items: fill-in, short-answer, and multiple-choice. Fill-in and short-answer items are best used in oral one-on-one tests, not in written tests. The reason is that there can be many possible correct answers to constructed-response items, and the answers can vary in level of detail. For example, if we asked a student to explain what happens in the circuit above when the button is pushed, the student might say that the bell rings. This answer is correct but is not at the level of detail we want. In an oral test, follow-up questions can be asked to probe the student's understanding.

In written tests, multiple-choice test items can be very effective in identifying misunderstandings, if the incorrect alternatives are constructed carefully. A special form of multiple-choice testing, called a "paper and pencil" simulation, can also be used when it is necessary to follow a student's logic through several diagnosis or prediction steps, particularly if there are several different correct ways of arriving at a solution. Constructing these test items is discussed below.

It is also possible to build computer-based simulations that allow the student to manipulate a system to make predictions or to perform tests and make decisions as in troubleshooting. An expert in computerized testing should be consulted.

Step 3. Construct Actual Test Items

Step 3A. Constructing Explanation Items. For explanation items, each of the cause-effect statements obtained in Step 1 becomes the stem for either a short-answer item (if this is an oral test) or for a written multiple-choice item. For multiple-choice items, the next step is to construct the distractors or incorrect alternatives. For short-answer items, the next step is to obtain information for scoring student responses. To do this, you should interview either experienced instructors or students just learning the material to determine what misconceptions are common. By "experienced," we mean that the instructor should be very familiar not just with the content but also with how students typically learn it and what mistakes they make or misconceptions they have. These interviews should be conducted by giving the short-answer cause and effect items and then asking the student to "think out loud" as the problem is solved, or asking the instructor to report what students typically do, both correctly and incorrectly. For example, in the circuit above, some students might think that current flows from battery 1 through to the bell when the relay is closed. This is not correct--battery 2 powers the bell when the relay contacts are closed--but it is a possible misconception about how relays work. If this is used as a distractor in the multiple-choice item, we will identify new students who have this misconception, and they can be remediated.

If alternate forms of the test are needed, use the paraphrasing techniques described in Chapter 3.

Step 3B. Constructing Prediction Items. For prediction items, each of the changes in boundary conditions or initial states identified in Step 1 becomes the stem for a short-answer or multiple-choice item. Typical prediction items are:

What would happen if (these new boundary conditions were true)...

What would be the effect on the operation of the system if (these changes in initial state occurred)...

After these questions, students can also be asked to justify their answers. In oral tests, this means the student is asked "why?" On written tests, multiple-choice explanation questions should be constructed. As stated in Step 1, this is done by building a new explanation for the new boundary conditions, and then treating this explanation as in Step 3A.

Step 3C. Constructing Diagnosis or Troubleshooting Items. For diagnosis or troubleshooting items, the student must "work backwards" from a given set of symptoms to determine the cause or causes. Symptoms occur when something does not work the way it should. In any system, there are usually several things that could go wrong to produce given symptoms. To construct diagnostic or troubleshooting test items, the things that typically break down and things that don't break down often but are critical to the system should be identified, and the resulting symptoms should be determined. These symptoms can then be used as the stem(s) for short-answer or multiple-choice items.

Distractors or alternatives for multiple-choice items should be constructed by analyzing each symptom to determine all the possible causes for it. Then, the causes that are not really responsible can be used as distractors. A typical item like this for our circuit example is:

> "Which of the following faults could cause the bell to ring even when the button is not being pressed? Choose all of the following which apply.
>
> A. A short across the contacts of the relay.
>
> B. A break in the wire from the switch to the relay coil.
>
> C. A short across the switch terminals.
>
> D. Using a latching switch or a latching relay."

Another way to test diagnosis is to construct a paper-and-pencil "simulated" troubleshooting exercise. This is done by building items (either fill-in or multiple-choice) that require the student to work backwards from symptoms to causes. In this case, since there may be several correct ways of troubleshooting, the whole sequence of answers is evaluated, not just the answer to any one question. Here is a set of items using our circuit example. (Notice that this example goes beyond the cause and effect analysis done in Step 1. Building this type of test requires additional information about troubleshooting techniques.)

> "Assume that the bell in the circuit above fails to ring when the button is pushed. What is the first test you would perform to begin isolating the fault?"
>
> Will you perform this test with the battery in the circuit or will you remove it? (That is, should the circuit be powered or not?)"

"Circle the points in the schematic above where you would place your probes to perform this test."

"What value would be observed if the circuit were operating normally?"

"What value would you observe if the switch contacts were dirty and failing to complete the circuit?"

(etc.)

Each of the questions above could be turned into a multiple-choice question, if the alternatives are carefully chosen. The alternatives should include realistic tests to perform, and the al'ernative values in later questions should include the values that should be obtained from each of the tests. Here is the same example as above but in multiple-choice format:

"Assume that the bell in the circuit above fails to ring when the button is pushed. Which one of the following tests would you perform first?

A. A voltage check of battery 1.

B. A continuity check of the entire bell circuit from the relay contacts.

C. Hold the relay contacts closed and see if the bell rings."

(etc.)

"Will you perform this test with the battery in the circuit, or will you remove it?

A. In B. Out"

"Give the letters of the points on the schematic where you would place your test probes? (Note: various test points on the schematic have been labeled.)"

"What should happen or what value should be observed when you perform this test if the circuit were operating normally?

A. Voltmeter should read 1.5 volts.

B. Ohmmeter should read 0 Ohms.

C. Bell should ring.

D. Bell should NOT ring."

(etc.)

Step 4. Develop Standards and Instructions for Scoring and Diagnosis

Step 4A. Setting Standards. Setting standards for USE-PRINCIPLE tasks is similar to setting standards for USE-CATEGORY tasks. Most tasks will require complete accuracy. However, there are some tasks that cannot be performed perfectly, even by job experts. For these tasks, the standards will be determined by the nature of the job. See Step 4A for USE-CATEGORY tasks.

Step 4B. Developing Instructions for Scoring and Diagnosis. See Step 4B for USE-CATEGORY tasks.

Step 5. Develop REMEMBER-level Test Items if Necessary

If REMEMBER-level objectives will be tested during the USE-PRINCIPLE test, develop the necessary test items using the procedures described in Chapter 3.

REMEMBER-level test items may be given orally or on a written test and may be administered during or prior to the USE-PRINCIPLE test. REMEMBER-level test items should not be given during the USE-PRINCIPLE test if they interfere with the PRINCIPLE task. This could occur if the PRINCIPLE task had to be performed in a set amount of time.

REMEMBER-level items given during a USE-PRINCIPLE test should test memory for symbol names, the decisions, predictions, and explanations relevant to the principle, and definitions of unfamiliar words and technical terms.

EXAMPLE: Designing Test Items for a USE-PRINCIPLE Objective

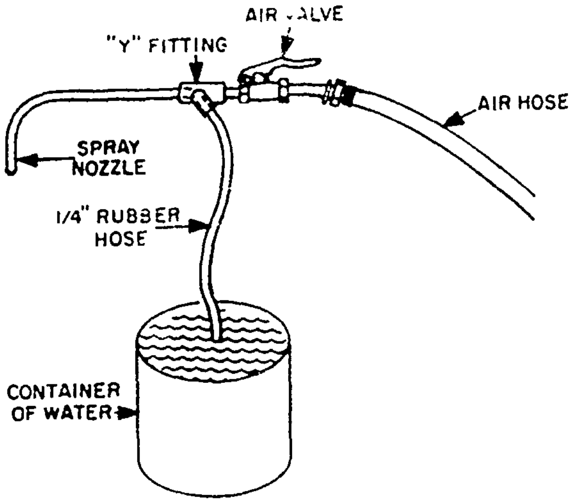
The following example is presented to illustrate the methods for developing test items for USE-PRINCIPLE objectives. For this purpose, the example was designed to be very simple. Thus, while it is a USE-PRINCIPLE task, it is simpler than most USE-PRINCIPLE tasks encountered in test design and development.

Here is a USE-PRINCIPLE objective:

"Given a diagram of a water spray torch, used for reducing cooling time in welding, the student will explain how the torch operates, predict the effect on the torch's operation of a clogged or perforated 1/4-inch rubber hose or a malfunctioning air valve, and diagnose the probable reasons for lack of water in the air stream or lack of flow."

Step 1. Analyze the objective to determine test item domain.

Here is a diagram of the water spray torch and an explanation of how it works.



Explanation of Water Spray Torch Operation:

The water spray torch (shown above) is used to reduce cooling time in some welding jobs. The torch consists of a nozzle, a "Y" fitting, a control valve, and a 1/4-inch rubber hose. The air hose is connected to the ship's compressed air and the rubber hose is run to a pail of water, as shown above. When the control valve is opened, the rush of air past the orifice in the "Y" fitting draws some water into the air stream, creating an atomized spray. When the spray strikes the hot plate, it turns into steam. A substantial amount of heat is absorbed in converting the atomized spray to steam. One pound of water will absorb 142

BTUs in vaporizing into steam. Because of the heat requirements for vaporization, the cooling is rapid, even with the use of only a small amount of water. Since all of the sprayed water is vaporized, the work will remain dry.

The first step in the analysis is to develop explanation, prediction, and diagnosis item types for the tasks required by the objective. These types of items are discussed below.

1. Explanation item types. Explanation item types are developed by specifying all the cause-effect relationships as single statements. The statements given below were developed by revising the explanation given above:

a. Compressed air flows through the air hose and air valve when the valve is opened.

b. The compressed air flows passed the "Y" fitting and creates a vacuum in the 1/4-inch rubber hose.

c. The vacuum created by the flowing compressed air causes water from the container to be drawn into the "Y" fitting.

d. The water and compressed air combine to form an atomized spray which flows out of the spray nozzle.

2. Prediction item types. Prediction item types are developed by giving new boundary conditions or initial states and then by developing a new set of cause-effect statements that explain how the system works with the new conditions. The following statements were developed based on the new boundary conditions specified in the objective.

If the air valve fails to open, the explanation would be:

a. Compressed air does not flow through the air valve.

b. No air flows passed the "Y" fitting and therefore, no vacuum is created in the 1/4-inch hose.

c. No air or water flows out of the spray nozzle.

If the 1/4-inch rubber hose is clogged, the explanation would be:

a. Compressed air flows through the air valve when the valve is opened.

b. The compressed air flows passed the "Y" fitting and creates a vacuum in the 1/4-inch rubber hose.

c. The clog in the rubber hose prevents water from the container from being drawn into the "Y" fitting.

d. Only air flows out of the spray nozzle.

(Note that the explanation for a perforated hose is similar.)

3. Diagnosis item types. Diagnosis items require the student to start from symptoms and work backwards through the explanation to identify possible causes. The objective asks the student to diagnose the probable reasons for lack of water in the air stream. This could occur for several reasons including (a) a clogged rubber hose, (b) a clog in the "Y" fitting, (c) a hole in the rubber hose that would prevent a vacuum from forming, or (d) an empty water container. It is important to realize that, for this simple example, diagnosis is not difficult. However, it is often necessary to consult an subject matter expert to determine possible causes for symptoms because most systems are more complicated than the one in this example.

Step 2. Determine item formats.

The objective does not specify item formats, so for this example we will assume that the student is being tested orally in a laboratory or job situation. Thus, the best type of item is short-answer. If a written test were required, fill-in items could be used. Multiple-choice items are not the best choice because the job the student will perform will not have selected-response situations. For example, hoses will not have four possible faults listed on them.

Step 3. Construct actual test items.

Step 3A. Explanation items. For explanation items, each of the statements from Step 1 becomes the stem for a short answer item or can have words deleted to be a fill-in item. For example:

Short-answer: "What happens when the compressed air flows passed the 'Y' fitting?"

Fill-in: "The compressed air flows passed the 'Y' fitting and creates a ___________ in the 1/4-inch rubber hose."

If multiple-choice items must be used, experienced instructors, or students learning the material should be interviewed to determine common misconceptions. These miconceptions should be used as distractors.

Step 3B. Prediction items. Constructing prediction items is similar to constructing explanation items. The statements developed from changing the boundary conditions in Step 1 become the stem for short-answer items or can have words deleted to be fill-in items. For example:

Short-answer: "What flows out the spray nozzle when the rubber hose is clogged?"

Fill-in: "A clog in the rubber hose prevents ________ from being drawn into the 'Y' fitting."

Step 3C. Diagnosis items. For diagnosis items, the student is given a symptom or set of symptoms and must "work backwards" to determine the probable cause or causes. For this objective, a short-answer oral question would be most appropriate. The correct answers would be the reasons identified in Step 1. For example:

Short-answer: "Tell me all the probable reasons for lack of water in the air stream?"

Step 4. Develop Standards and Instructions for Scoring and Diagnosis.

Step 4A. Standards. The implied standard in the objective is "100% correct," and there is not reason to set a lower standard for a task this simple.

Step 4B. Instructions for Scoring and Diagnosis. For oral short-answer explanation, prediction, and diagnosis questions, the student should be told as much as possible about the form and content of an acceptable answer.

Instructions for scoring all types of oral questions should list the critical points or answers that the student should give. Common errors should be listed, so that they can be diagnosed and corrected.

Instructions for scoring fill-in question should list all correct answers and synonyms.

For a simple task like this example, diagnosing why errors are made is not difficult and extensive instructions or diagnostic plans are not required. However, for more complex tasks, more detailed plans and instructions may be needed.

Finally, instructions for scoring and diagnosis should include remedial actions to be taken for incorrect answers. For the present task, the best type of remediation would be oral remediation by the instructor or supervisor.

Step 5. Develop REMEMBER-level Test Items if Necessary

No REMEMBER-Level objectives are tested in this example.

CHAPTER 6

TEST TRYOUT AND TEST ITEM ANALYSIS

Introduction

Once the test items have been written and assembled into tests, it is necessary to try them out on real students to make sure that they are not ambiguous and that they test the objectives accurately. The tryout procedure is also necessary to detect inconsistencies in the instructional materials. It should be emphasized that the tryout is a tryout of both the test items and the instruction. It is very possible to have a good test item that everyone misses because the instruction is poor.

Some of the procedures that will be described in this section are statistical. However, it is important to realize that statistical procedures can only be used in criterion-referenced testing to identify or "flag" test items that may be poor items. Further investigation will be necessary to determine whether the test item is flawed or the instruction is deficient. Therefore, in addition to performing statistical tests on the items, items must be thoroughly reviewed to make sure they have no obvious technical flaws. Procedures for doing this type of review will also be described in this chapter.

Finally, the statistics described in this chapter are very simple and can be calculated with a hand calculator. There are more sophisticated techniques for item analysis; however, these techniques require the use of a computer. If you have computer facilities available for item analysis, you should contact the Chief of Naval Education and Training, Code N-9, for appropriate analysis techniques.

Test Item Analysis Procedures

Step 1. Select a Sample of Students for Test Tryout

Try to obtain as many students as possible for the tryout. Your decisions about the instruction and test items will be more accurate if you have 15 students in your sample than if you have two. A good rule of thumb is to use about half the number of students that would be in an average class during the actual course. However, the more students you can get, the more reliable and accurate your analysis will be. The students selected for test tryout must meet at least two criteria.

First, they should be representative of the population of students that the test is intended for. You should not try out test items for steam propulsion on students from the cooks and bakers school. These items should be tried out on students from the propulsion engineering school.

Second, for most Navy introductory courses, half the students in the sample should have gone through the instructional materials and half should not be instructed. It is assumed that the uninstructed students will know very little about the content and that the instructed students will have much more knowledge. The only time it is unnecessary to have uninstructed students is when you can be sure that the uninstructed group will have NO knowledge of the content. In this case, you can assume that their scores on the test items will be zero. This situation occurs more frequently in advanced courses.

If students are in short supply,it is possible to give the instructed students the test twice--once before they go through the course and once after they complete the course. This is called a pretest-posttest comparison. The problem with this method is that, if the instruction is given immediately after the pretest, the pretest could influence the way in which students learn the materials and as a result affect posttest performance. In a tryout situation, it is often necessary to give the instruction and test in a short period of time, so the pretest-posttest method is not usually appropriate.

Step 2. Administer Test Items

The test items should be given to the sample of students. The tryout should be administered in a standardized fashion, just as if you were giving the test in the course. The following are some guidelines that should be observed, if possible, during the tryout:

1. If possible, have someone else administer the tryout, so that you can be free to observe the process and note problems. It is best to have more than one observer and more than one administrator.

2. Individuals in the sample should be informed that they are serving in a tryout to help develop a test. They should be asked to make notes of confusing or ambiguous items, and of anything they don't understand.

3. Use the instructions developed previously and administer items as specified. The tryout is also used to evaluate the instructions; lack of clarity or ambiguity should be noted by individuals in the tryout sample.

4. The test conditions should be the same for the tryout as they will be in the final version of the test. Do not try to short-cut the specified conditions as this will affect your tryout results. For example, if items require the use of a VT52 terminal, do not use a 1620D. If a test item calls for administration out of doors, give it outdoors, not inside.

5. Test standards should be the same in the tryout as in the final version of the test. You must be careful to score the items for the people in the tryout exactly as you will for the final version of the test.

Step 3. Analyze Results of Tryout

The principal use of item analysis data in criterion-referenced testing is to detect bad items. It is important to realize that such data, no matter how carefully analyzed, do not provide an absolute indication that an item is or is not flawed. Also, if an item is flawed, the data cannot tell the test developer exactly how to correct the flaw. What the data can do is "flag" a potentially flawed item and usually suggest the nature of the problem and/or the part of the item that is flawed. The procedures in this step are concerned with how to perform item analysis and interpret the results.

There are several simple statistical procedures that can be used to analyze tryout data. It is recommended that at least two different statistics be calculated for each item. The statistical test used will depend on the type of test item and whether or not you use the pretest-posttest method. The following sections explain these different types of procedures for selected-response items, constructed-response items, and performance and essay items, respectively.

Analysis Procedures for Selected-response Items

After the test items are administered, the following chart should be constructed for each item. We have used a multiple-choice item in the example, but the same procedure applies to matching and true-false items. All that varies is the number of alternatives. In this example, we are assuming that a multiple-choice test was given to 16 students, 8 who received the instruction and 8 who were not instructed. The chart displays the data for one item. Note that the same procedure could be applied if the tryout was given by the pretest-posttest method.

Sample Data for One Multiple-choice Item

	Group		
Alternative Answer	Uninstructed or Pretest	Instructed or Posttest	Total
a	3	1	4
b*	1	6	7
c	2	0	2
d	1	1	2
Omit	1	0	1
Total	8	8	16

Alternative b is the correct alternative. The chart shows the number of students in each group who responded to each of the four alternatives.

Two Statistical Tests

The P Statistic

The first statistic to be computed is the proportion of students who choose each alternative (or omitted an item). This statistic, which is called P, is computed by dividing the number of students who responded to each alternative by the total number of students in the sample. In the present example you would divide each number in the total column by 16. This would give you the following results:

Alternative a = 4/16 = .25
b = 7/16 = .44
c = 2/16 = .13
d = 2/16 = .13
omit = 1/16 = .06

The P statistic tells you how easy the item was for the students in your sample. In the present example, half the students have not been given the instruction, and you would expect that these students would not answer the item correctly. In this situation, the P statistic should be between .35 and .65. If it is lower than .35, this means that many of the students who have been through the instruction are not answering correctly. If it is higher than .65, then many students who have not had the instruction are answering correctly.

If you were in a situation where you did not have an uninstructed group, you would expect P to be very high. If the P value in this case is less the 1.0, students who have been instructed are missing the item. In an ideal world, you would not want this to occur. In the real world, if the P value is below .80, the item should be reviewed.

The P statistic should be interpreted using the following guidelines:

1. Look at the P value for the correct alternative. The item may be flawed if P is considerably out of line with a value one might expect. In our example, if P is larger than .65 or smaller than .35, the item should be reviewed using the methods for reviewing test items listed in this chapter. If the P value is larger than .65, the objective should also be reviewed to determine whether it should be included in the curriculum. Remember that a P value higher than .65 means that many of the uninstructed students are getting the items correct. When this occurs, it may not be necessary to teach or test the information covered by that item. If all the subjects in the sample have been given the instruction, the item should be reviewed if P for the correct alternative is less than .80.

2. Look at the P values for the distractors. If the P value for any distractor is more than the P value for the correct alternative, the distractor should be examined to see if it could be considered, reasonably, as a correct answer. If so, one of three problems probably exist: (a) the correct answer was not written correctly, (b) the item has two or more correct answers, or (c) the item is ambiguous. In these cases, revision is necessary. Another possibility, of course, is that the distractor is fine, and that students are being misled by a fault in the instruction. In this case, it is necessary to review the instruction for that objective and make sure that students are not developing misconceptions.

3. The P values for the distractors should be examined to determine if any of the P values are very small.. If this occurs, you should consider eliminating the alternative and replacing it with some other incorrect alternative, provided that doing so does not change the intended nature of the item. You should NOT change distractors if they have been constructed systematically to be diagnostic, even if the P value is low. Note that, if the quality of the instruction is good and the item is not hard, it is likely that one or more of the distractors will be chosen infrequently.

The D Statistic. The second statistic to be computed is the proportion of students in the instructed group who chose an alternative (or omitted an item), minus the proportion of students in the uninstructed group who chose an alternative (or omitted an item). This statistic is called D--the Discrimination Statistic. D is computed by (1) dividing the number of instructed students who responded to each alternative by the total number of instructed students, (2) dividing the number of uninstructed students who responded to each alternative by the total number of uninstructed students, and (3) subtracting the second result from the first. For the example, the results are as follows:

Instructed - Proportion Choosing Each Alternative
Alternative a = 1/8 = .13
b = 6/8 = .75
c = 0/8 = .00
d = 1/8 = .13
omit = 0/8 = .00

Uninstructed - Proportion Choosing Each Alternative
Alternative a = 3/8 = .37
b = 1/8 = .13
c = 2/8 = .25
d = 1/8 = .13
omit = 1/8 = .13

D Statistic for Each Alternative
Alternative a = .13 - .37 = -.24
b = .75 - .13 = .62
c = .00 - .13 = -.13
d = .13 - .13 = .00
omit = .00 - .13 = -.13

The D statistic tells you how well the item discriminates between students who have been instructed in the subject matter and students who have not been instructed. D values range from -1.0 to +1.0. If more instructed students than uninstructed students choose an alternative, D for that alternative will be positive; that is, it will be between 0 and +1.0. If, on the other hand, more uninstructed students than instructed students choose an alternative, D will be negative; that is, it will be between -1.0 and 0. The D statistic should be interpreted using the following guidelines:

1. Look at the D value for the correct alternative. It is very unlikely that a good item would have a value for D that is negative, because that would mean that a greater proportion of the uninstructed students got the item correct than the instructed group. Therefore, if D is negative, the item should be reviewed, checking especially to see that the item was scored correctly, that it is unambiguous, and that the indicated correct

answer in really correct. If D for the correct alternative is 0 or just slightly positive, this means that uninstructed students are doing almost as well as instructed students. In this case, the items should also be reviewed. In addition, the objective should be reviewed to determine whether or not it should be included in the curriculum. The test item review procedures are described later in this chapter. If the item is O.K., the instruction for that item should be reviewed.

2. Look at the values of D for the distractors. If any of them are positive, check the item to see if it is ambiguous or if the distractor could possibly be a correct answer. Here again, it is possible for the distractors to be fine and the instruction to be at fault.

3. If either P or D for the "omits" is positive, examine the item for ambiguities. It is assumed that students are not being penalized for guessing. Therefore, there is no reason for a student to skip an item.

In the example presented in the chart, the item is O.K. for both the P and D statistics.

Analysis Procedures for Constructed-response Items

The analysis procedures for constructed-response items are very similar to the procedures for selected-response items. The procedures described in this section should only be used for fill-in or short-answer questions. Analysis techniques for essay-type questions will be discussed in the section on analysis of performance tests.

First, a chart, similar to the chart for selected-response items, should be constructed. In the example below, we have administered the test to 10 uninstructed students and 10 instructed students. The chart summarizes the data for items 1 through 5 on the test. The numbers in the columns indicate, for each group, the number of students out of 10 who answered the item correctly. Note that the same procedures could be applied if the tryout was given by the pretest-posttest method.

Sample Data for Five Constructed-response Items

	Group	
Item Number	Uninstructed or Pretest	Instructed or Posttest
1	0	8
2	3	6
3	6	7
4	10	10
5	4	7

In the chart above, an item omitted by a student was scored as incorrect.

Two Statistical Tests

The K Statistic. The first statistic to be computed is the proportion of students in the instructed group who answered an item correctly minus the proportion of students in the uninstructed group who answered it correctly. This statistic is called K. K is computed by (1) dividing the number of instructed students who responded correctly to each item by the total number of instructed students, (2) dividing the number of uninstructed students who responded correctly to each item by the total number of uninstructed students, and (3) subtracting the second result from the first. For the example, the results are as follows:

Instructed - Proportion Correct on Each Item

Item	1	=	8/10	=	.80
	2	=	6/10	=	.60
	3	=	7/10	=	.70
	4	=	10/10	=	1.00
	5	=	7/10	=	.70

Uninstructed - Proportion Correct on Each Item

Item	1	=	0/10	=	.00
	2	=	3/10	=	.30
	3	=	6/10	=	.60
	4	=	10/10	=	1.00
	5	=	4/10	=	.40

K Statistic for Each Item

Item	1	=	.80	-	.00	=	.80
	2	=	.60	-	.30	=	.30
	3	=	.70	-	.60	=	.10
	4	=	1.00	-	1.00	=	.00
	5	=	.70	-	.40	=	.30

The K statistic tells you how well the item discriminates between students who have been instructed in the subject matter and students who have not been instructed. K values range from -1.0 to +1.0. If more instructed than uninstructed students answer the item correctly, K will be positive. If, on the other hand, more uninstructed students than instructed students answer an item correctly, K will be negative. The K statistic should be interpreted using the following guidelines:

1. The K value for each item should be positive. If there are any items with a negative K, the item should be reviewed for ambiguity and/or the instruction for that item should be checked. Item review procedures are described later in this chapter.

2. If the K value for an item is between .00 and .30, this means that about as many uninstructed students are getting the item correct as instructed students. This can happen if the item is so easy that almost everyone gets it right (see item 4 in the example) or if the item is so difficult that many students miss it (see items 2, 3 and 5 in the example). In either case, the item should be reviewed. If the item is too easy, it should either be revised or the objective should be reviewed to determine if the item should be deleted from the curriculum. If it is too hard, it should be reviewed for ambiguity and/or the instruction should be checked.

The A Statistic. The A statistic is the proportion of instructed students correct on each item and has already been calculated as part of the K statistic (see the first column in the results table above). This statistic should be used in conjunction with the second check of the K statistic described above. If the value of A is below .70, the item should be reviewed for difficulty and/or the instruction should be checked. Item review procedures are included later in this chapter.

Do You Have Enough Items or Too Many?

The statistical methods discussed above are meant to "flag" individual selected- and constructed-response items that may need revision. There is, however, another issue--determining whether you have enough items or too many--that can be dealt with statistically. Procedures for doing this are described in this section. It is important to note that the statistics described in the previous sections should be done on selected and constructed-response test items before the procedures in this section are applied to them.

The problem of determining how many items are needed only occurs with some tasks. For REMEMBER-level tasks and USE-PROCEDURE tasks, there is essentially only one or a small number of possible items, so all of them are tested. However, tasks that require transfer define lots of items. As we mentioned in the previous chapter, the problem for USE-CATEGORY, USE-RULE, and USE-PRINCIPLE tasks is to determine how many items must be given to students so that their performance can be assessed. For USE-CATEGORY tasks, we built several items that test the presence or absence of one critical characteristic. For USE-RULE tasks, we built several items for each "path" in the rule, and for USE-PRINCIPLE tasks, we built several different prediction or troubleshooting items. In the previous chapter, we postponed the problem of determining how many items must actually be administered until after tryouts were completed. So, here we are.

Some examples might clarify the issue. In our call signs task, we could have built lots of items for each type of call sign. For example, NKVD, NABC, NPSR, items are all correct International Ship call signs. How many of these must the student classify? In our fowl task, we could have built lots of items for each path. For example, we could have had the student calculate cooking times for a 2 lb., 2.5 lb., 3 lb., 3.1 lb., 3.45 lb, etc. stuffed fowl. (All of these use the same path.) Again, how many of these must a student calculate?

Because we have more than one possible item, we also have a further problem. Suppose we give students more than one item of each type. But, suppose their performance is not uniform; that is, suppose for the same type of item, they get some items right and others wrong (instead of all right or all wrong). Do we need more items, or do we need to rethink the analysis that led to the items in the first place?

To answer these questions on the basis of tryouts, it is necessary to administer two items of each type, or category, or path, to tryout students. Then we look at patterns of responses on both items of each type. We will arrange the results in the following table:

		Second Item	
		correct	incorrect
First Item	correct	Proportion of students who got both items right	Proportion of students who got first right and second wrong
	incorrect	Proportion of students who got first wrong and second right	Proportion of students who got both items wrong

(Note: Proportions are calculated by dividing the number of students in a box by the total number of students in all four boxes. You will have already calculated these proportions for the statistics described previously.)

Suppose we gave two items to 20 students who had received instruction. Ideally we would expect all students to get both items right. The results would be:

		Second Item	
		correct	incorrect
First Item	correct	20/20 = 1.0	0
	incorrect	0	0

For uninstructed students, we would expect all students to miss both items:

		Second Item	
		correct	incorrect
First Item	correct	0	0
	incorrect	0	20/20 = 1.0

In a real situation, our results would not be so clean. The most desirable situation is to have students in either the "correct-correct" box or the "incorrect-incorrect" box. This means students either know the content or not. On the other hand, if some students are getting one item correct and the other incorrect, something is wrong somewhere. In other words, our items are O.K. if students are getting them both right or both wrong, but we have a problem otherwise. This situation is illustrated below:

		Second Item	
		correct	incorrect
First Item	correct	GOOD	BAD
	incorrect	BAD	GOOD

Suppose 12 of our 20 students got both items right, and the other 8 got both items wrong. The results would be:

		Second Item	
		correct	incorrect
First Item	correct	12/20 = .60	0
	incorrect	0	8/20 = .40

From these results, we can conclude that it is NOT necessary to administer both items on the actual test. The reason is that the two items are "perfectly" correlated; that is, if a student gets one item right (or wrong), the other must be right (or wrong) too.

Let's consider some other possible results. Suppose a very few students missed either the first or second item but got the other correct, as in the table below:

		Second Item	
		correct	incorrect
First Item	correct	11/20 = .55	1/20 = .05
	incorrect	0/20 = 0	8/20 = .40

These results are not too bad. Only one student is giving BAD results, and this is probably due to a careless error. In general, if the proportion in the BAD boxes is less than about .05, we can still use just one item of each type instead of both. If the proportion in the BAD boxes is more than .05 but no larger than .20, then both items should be used on the test, and instructors should be careful to diagnose student errors. If the proportion in the BAD boxes is more than .20, we have more serious problems. Either one of the items is flawed, or the original analysis is in error. To distinguish these possibilities, we look at the two BAD boxes. If most students are in one of the BAD boxes but not the other, one or the other item is flawed. Suppose we got the following results:

		Second Item	
		correct	incorrect
First Item	correct	7/20 = .35	10/20 = .50
	incorrect	0	3/20 = .15

These results mean either that the first item has hints or cues to the correct answer, or that the second item is ambiguous or misleading in some way. The other statistical methods discussed earlier in this chapter can be used to identify the bad item.

If lots of students are spread more or less evenly over both BAD boxes, then the original analysis is in error. Suppose we got the following results:

		Second Item	
		correct	incorrect
First Item	correct	2/20 = .10	8/20 = .40
	incorrect	6/20 = .30	4/20 = .20

This means that the test items are not working the way they were designed. For CATEGORIES, this might mean that some critical characteristic has been overlooked, or that students are being mislead by some irrelevant characteristic. For RULES, some decisions (and steps) have been left out of the flow chart, and so what seemed to be one path is really two or more. For PRINCIPLES, predictions or diagnoses are not being made in quite the same way in the two items. In any case, the original analysis and design of the items must be carefully checked and revised if necessary. If time constraints or the complexity of the subject matter make it difficult or impossible to check the analysis thoroughly, more test items should be given. A good rule to follow is that "if you are confident that the analysis is accurate, you can use a minimum number of test items; on the other hand, if your confidence in the analysis is low and your tryout data are not good, more than a minimum number of test items are required."

Analysis Procedures for Performance and Essay Items

Performance and essay items require a different type of analysis than constructed- and selected-response items, because performance and essay items are almost always scored with a checklist or rating scale. The first analysis that should be done is to check the reliability of the checklist or rating scale. Once you are confident that the checklists or rating scales are reliable, the student reponses can be analyzed to see if there are any items or parts of items that need to be reviewed. The following sections describe the types of errors that raters can make and procedures checking the reliability of checklists and rating scales for performance and essay items.

Types of Rating Errors

One problem with rating scales is that different raters often make different judgments about the same performance. These differences or rating errors can be classified into four categories:

1. Error of Standards. Errors are sometimes made because of differences in different raters' standards. If rating is done without any specified standards, there may be as many different standards as there are observers. This is why it is important that rating scales be "anchored" with descriptions of the behaviors for each value on the rating scale. The more complete these descriptions, the better the interrater agreement.

2. Error of Halo. A rater's ratings may be biased because he allows his general impression of an individual to influence his judgment. This results in a shift of the rating and is known as a "halo" effect. If a rater is favorably impressed, the shift is toward the high end of the scale. If the rater is unfavorably impressed, the shift is toward the low end. This type of error frequently goes undetected unless it is extreme. It is therefore a difficult error to overcome. Error of halo is reduced by reminding each rater that he is judging specific performances and should NOT take into consideration his overall impression of a student.

3. Logical Error. A logical error may occur when a rater uses a series of rating scales. When a rater tends to give similar ratings on scales that are not necessarily related, he is making a logical error. The way to minimize logical errors is to make clear the distinctions among different performances or aspects of a product that are to be measured. Again, behavioral "anchors" help.

4. Error of Central Tendency. An error of central tendency is demonstrated when different raters tend to rate most students near the middle of a scale. If, for example, a scale has seven points and you get a large number of "4s" from the raters, they may be making this error. One way to counter this is to use scales with an even number of points (so there is no middle point). Also, behavioral "anchors" again help.

Determining Reliability of Checklists and Rating Scales

Rating Scales. Rating scales are used in items that involve decisions more complicated than "yes-no" or "go-no-go." It is important that different raters use the scale in the same way. To determine how well different raters agree, you should construct a chart similar to the one below. The chart should show the score that each rater gave to each student on each item. In the example below, three raters rated five students on five items. The rating scale for each item was 1 to 5.

Sample Rating Scale Data for Five Items

	Student 1			Student 2			Student 3			Student 4			Student 5		
Item Number	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
1	5	5	5	3	3	3	4	4	4	2	2	2	1	1	1
2	5	4	4	4	4	4	3	4	3	1	2	2	2	3	2
3	5	4	5	4	3	3	3	3	3	3	2	2	4	4	4
4	3	5	2	3	1	4	2	4	3	1	2	4	3	2	1
5	4	4	3	3	2	3	4	3	4	3	2	2	3	3	3

R1=Rater 1, R2=Rater 2, R3=Rater 3.

By looking across a row, you can compare the scores given by the different raters to each student. In the example above, you can see that for item 1, there is perfect agreement among raters. For items 2, 3, and 5, there is some disagreement and for item 4, there is considerable disagreement. A good guideline is that, if the majority of raters agree and the raters disagreeing are only off by one point on the scale, the rating scale is reliable. If, however, there is no majority agreement or if raters differ by 2 or more points on the scale, a review is necessary. The rating scales should be checked to make sure that the "anchoring" statements are as clear as possible, and the instructions to scorers should be checked to make sure they are not misleading some of the raters. It is best to do this with the raters, because they can tell you what they thought they were doing.

Checklists. Checklists should be treated in the same way as rating scales, except that there will be only two possible scores, 1 or 0, yes or no, or go or no-go. Again, different raters should be compared with each other, as in the table above, to determine if there is substantial disagreement. If so, the checklists and instructions to scorers should be reviewed.

Essay Test Items. Since essay items are best scored using checklists or rating scales for major points in the answers, the procedures described above are applicable.

Methods for Reviewing Test Items

So far we have discussed statistical methods for "flagging" items that may be flawed. There are other, less formal, follow-up methods for reviewing items, which should be used to correct these flaws. These methods are discussed below.

1. Feedback from students. Feedback from individuals in the tryout can be extremely useful in identifying flaws. Interview as many students in the tryout as possible. Have them "walk through" their thinking as they responded to items. You should note difficulties with instructions or with particular items, time pressures, problems with equipment or facilities, misunderstandings of standards or scoring, and other points of confusion. It is best to conduct this review orally with individual students, because you can ask follow-up questions to pinpoint the source of problems.

2. Peer review. Another useful technique is to have experienced test developers review your items.

3. Review by test evaluators. Many Navy commands have a Curriculum and Instructional Standards Office (CISO), or evaluator, responsible for quality control. They will have their own procedures for review and revision of tests and their own sets of criteria that tests should meet. One set of test consistency and adequacy criteria, which are entirely consistent with this manual, appears in the Instructional Quality Inventory (IQI).

4. Review by subject matter experts. You should always obtain reviews of your test items from subject matter experts. They should be asked to check the items for technical accuracy and to note items that are confusing or misleading.

5. Review of practice items. If the practice items completed by instructed students are available, they can be used to help review test items. Since practice items should be similar or identical to the test items, performance on practice items can be compared to performance on related test items. If there are major differences between performance on practice and related test items, the items should be reviewed using the procedures described in this section. Some additional things to look for are inadequate instruction, long delays between initial training and testing (which could result in forgetting), practice items and test items that are inconsistent, and inappropriate sequencing of instruction, such that practice items occur before a proper instructional foundation has been laid.

REFERENCES

Anderson, R. C. How to construct achievement tests to assess comprehension. *Review of Educational Research*, 1972, *42*, 145-170.

Brennan, R. L. *Some statistical procedures for domain-referenced testing: A handbook for practioners* (NPRDC Tech. Note 81-6). San Diego: Navy Personnel Research and Development Center, February 1981. (Also published as ACT Technical Bulletin No. 38 by the Research and Development Division, The American College Testing Program, Iowa City, Iowa 52243).

*Brown, J. S., & Burton, R. R. Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science*, 1978, *2*, 155-192.

CNET. *Interservice procedures for Instructional Systems Development* (NAVEDTRA 106A). August 1975.

CNET. *Procedures for Instructional Systems Development* (NAVEDTRA 110A). September 1981

*Courseware, Inc. *Author training course* (NAVEDTRA 10003). 1978.

*de Kleer, J., & Brown, J. S. Mental models of physical mechanisms and their acquisition In J. R. Anderson (Ed.), *Cognitive skills and their acquisition*. Hillsdale, NJ: Lawrence Erlbaum, 1981.

*Durnin, J. H., & Scandura, J. M. An algorithmic approach to assessing behavior potential: Comparison with item form and hierarchical technologies. *Journal of Educational Psychology*, 1973, *64*, 262-272.

Ebel, R. L. *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall,1979.

Ellis, J. A., & Wulfeck, W. H., II. *The instructional quality inventory: IV. Job performance aid* NPRDC Spec. Rep. 79-5). San Diego: Navy Personnel Research and Development Center, November 1978 (AD-A083 928)

*References indicated by an asterisk are not cited directly in the handbook. However, most of the procedures and guidelines presented in the handbook are based on these references.

Ellis, J. A. & Wulfeck, W. H. Assuring objective-test consistency: A systematic procedure for constructing criterion-referenced tests (NPRDC Spec. Rep. 80-15). San Diego: Navy Personnel Research and Development Center, March 1980.

Ellis, J. A., Wulfeck, W. H., II, & Fredericks, P. S. The instructional quality inventory: II. User's manual (NPRDC Spec. Rep. 79-24). San Diego: Navy Personnel Research and Development Center, August 1979 (AD-A083 678)

Ellis, J. A., Wulfeck, W. H., II, & Fredericks, P. S. Workbook for testing in Navy schools (NPRDC Spec. Rep. 83-7). San Diego: Navy Personnel Research and Development Center, November 1982.

Fredericks, P. S. The instructional quality inventory: III. Training workbook (NPRDC Spec. Rep. 80-25). San Diego: Navy Personnel Research and Development Center, July 1980 (AD-A092 804)

Horabin, I. & Lewis, B. The Instructional Design Library, Vol. 2: Algorithms. Englewood Cliffs, NJ: Educational Technology Publications, 1978.

Lewis, B. Case Studies in the Use of Algorithms. Oxford, Pergamon Press Ltd, 1978.

*Mager, R. F. Measuring Instructional Intent or Got a Match? Belmont, CA., Fearon, 1973.

*Mallory, W. J., & Elliot, T. K. Measuring troubleshooting skills using hardware-free simulation (AFHRL-TR-78-47). Lowry AFB, CO: Technical Training Division, Air Force Human Resources Laboratory, December 1978.

*Markle, S. M. Designs for instructional designers. Champaign, IL: Stipes, 1978.

McGuire, C., Solomon, L., & Bashook, P. Construction and Use of Written Simulations. The Psychological Corporation, 1976.

* References indicated by an asterisk are not cited directly in the handbook. However, most of the procedures and guidelines presented in the handbook are based on these references.

*Merrill, M. D., & Tennyson, R. D. *Teaching concepts : An instructional design guide*. Englewood Cliffs, NJ: Educational Technology Publications, 1977.

*National Board for Respiratory Therapy, *Clinical simulation problem construction workbook: A guide for development of patient management problems in respiratory therapy*. Shawnee Mission, Kansas, 1978.

*Richards, R. E. *Principle learning* (unpublished manuscript), 1979.

Roid, G., & Haladyna, T. *Handbook of item writing for criterion-referenced tests* (NPRDC Tech. Note 80-8). San Diego: Navy Personnel Research and Development Center, February 1980.

*Smith, H. W., Frederickson, E. W., & Pearlstein, R. B. *Constructing diagnostic tests: Volume I. Theoretical and practical considerations*. Valencia, PA: Applied Science Associates, November 1979.

*Smith, H. W., Frederickson, E. W., & Pearlstein, R.B. *Constructing diagnostic tests: Volume II. Procedures and guidelines*. Valencia, PA: Applied Science Associates, November 1979.

*Swezey, R. W., & Pearlstein, R. B. *Developing criterion-referenced tests*. Valencia, PA: Applied Science Associates, 1974.

*Tiemann, P. W., & Markle, S. M. *Analyzing instructional content: A guide to instruction and evaluation*. Champaign, IL: Stipes, 1973.

Wulfeck, W. H., Ellis, J. A., Richards. R. E., Wood, N. D., & Merrill, M. D. *The instructional quality inventory: I. Introduction and overview* (NPRDC Spec. Rep. 79-3). San Diego: Navy Personnel Research and Development Center, November 1978. (AD-A062 493)

---

* References indicated by an asterisk are not cited directly in the handbook. However, most of the procedures and guidelines presented in the handbook are based on these references.

DISTRIBUTION LIST

Chief of Naval Education and Training (02), (N-2), (N-5), (N-9)
Chief of Naval Technical Training (016)
Commander Training Command, U.S. Atlantic Fleet
Commander Training Command, U.S. Pacific Fleet
Commanding Officer, Fleet Anti-Submarine Warfare Training Center, Pacific
Commanding Officer, Fleet Combat Training Center, Atlantic
Commanding Officer, Fleet Combat Training Center, Pacific
Commanding Officer, Fleet Training Center, San Diego
Commanding Officer, Naval Damage Control Training Center
Commanding Officer, Naval Education and Training Program Development Center (Technical Library) (2)
Commanding Officer, Naval Education and Training Support Center, Pacific
Commanding Officer, Naval Health Sciences Education and Training Command
Commanding Officer, Naval Regional Medical Center, Portsmouth (ATTN: Medical Library)
Commanding Officer, Naval Technical Training Center, Corry Station (Code 101B)
Commanding Officer, Naval Training Equipment Center (Technical Library)
Commanding Officer, Recruit Training Command (Academic Training Division)
Director, Defense Activity for Non-Traditional Education Support
Director, Management Information and Instructional Activity Branch Office, Memphis
Director, Naval Education and Training Program Development Center Detachment, Great Lakes
Director, Naval Education and Training Program Development Center Detachment, Memphis
Director, Training Analysis and Evaluation Group (TAEG)
President, Naval War College (Code E114)
Superintendent, Naval Postgraduate School
Commander, Army Research Institute for the Behavioral and Social Sciences, Alexandria (PERI-ASL)
Chief, Army Research Institute Field Unit--USAREUR (Library)
Chief, Army Research Institute Field Unit, Fort Harrison
Commander, Air Force Human Resources Laboratory, Brooks Air Force Base (Scientific and Technical Information Office)
Commander, Air Force Human Resources Laboratory, Lowry Air Force Base (Technical Training Branch)
Commander, Air Force Human Resources Laboratory, Williams Air Force Base (AFHRL/OT)
Commander, Air Force Human Resources Laboratory, Wright-Patterson Air Force Base (AFHRL/LR)
Commander, 314 Combat Support Group, Little Rock Air Force Base (Career Progression Section)
Commandant Coast Guard Headquarters
Commanding Officer, U.S. Coast Guard Institute
Commanding Officer, U.S. Coast Guard Research and Development Center, Avery Point
Commanding Officer, U.S. Coast Guard Training Center, Alameda
Superintendent, U.S. Coast Guard Academy
Defense Technical Information Center (DDA) (12)